UNIVERSIDAD PABLO DE OLAVIDE

Programa de Doctorado en Biotecnología, Ingeniería y Tecnología Química (R.D.99/2011)

TESIS DOCTORAL

Integrative Cell Biology

Biología Celular Integrativa

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Nicola Bordin

Director

Damien Paul Devos

Sevilla, 2018

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.“

Marie Skłodowska Curie

D. DAMIEN PAUL DEVOS, DOCTOR EN BIOLOGÍA POR LA UNIVERSIDAD DE NAMUR (BELGICA), Y CIENTIFICO TITULAR DEL CENTRO SUPERIOR DE INVESTIGACIONES CIENTIFICAS (CSIC) EN EL CENTRO ANDALUZ DE BIOLOGIA DEL DESAROLLO (CABD)

INFORMA

Que **DON NICOLA BORDIN**, Licenciado en Biotecnología, ha realizado bajo su dirección el trabajo titulado **"Integrative Cell Biology"**, y que a su juicio reúne los méritos suficientes para optar al grado de Doctor en Ciencias.

Y para que conste,

firmo el presente en Sevilla, a de de dos mil dieciocho.

Fdo.: Damien Paul Devos

Acknowledgments

These four years have been a heck of a ride. It's been challenging and stressful, but the possibilities, what I learned and the amazing people that helped, supported and blessed me with their friendship made it worthy and without a doubt the best period of my life so far.

This chapter is for you, for everyone that I met and bonded with during these years, and that's a lot of people. I'll try to thank everyone, but if you're not included bear with me, you meant something to me.

I'd like to thank first Damien, my friend, mentor and more than often my main source of headaches. Jokes aside, thank you for giving me the opportunity to work with you and with great collaborators. Thanks for introducing me to the magical world of the PVCs, teaching me how to write a manuscript, how to be thorough and properly deal with the University administration. I loved our discussions in front of a proper beer, the barbeques and pisco sours at your home, where I always felt welcomed by Miriam and your family. I'd like to thank you for making me what I am now. For allowing me to manage the collaborations and almost always have the final say on my own work. Thanks for teaching me first and foremost a thing that isn't usually possible during grad school, independence. I will treasure all these lessons for my postdocs and future years in academia. Good luck finding your next coffee maker!

I'd like to thank also my other mentors and collaborators, especially Sean, Olga and Lise. You've entrusted me with your time, data and patience. You've always been absolutely supportive and treated me as an equal, listening to my crazy ideas without dismissing them straight away. My gratitude goes also to my mates in Sydney, for the interesting chats over lunch at the cheap thai and fancy thai. Chris, Benedetta and Jenny, thanks for making me feel home even 18 thousand kilometers away from it.

Un agradecimiento especial a mis compis de laboratorio por soportar mi ropa de bicicleta, los hábitos raros, mi andaluz no perfecto, por ayudarme a traducir el correo ocasional al "Programa de doctorado internacional UPO" y por estar allí cada vez que necesitaba hablar de algo. Carlos, has sido más que un colega. Has sido un amigo, un entusiasta de la drum-n-bass y siempre me ha sugerido buena música para escuchar mientras trabajo. Muchas gracias también a Elena, por obligarme a mejorar

mi español, mis cursivos erróneos al nombrar los PVC y por ayudarme en muchas ocasiones. Siempre has estado ahí para mí cuando necesitaba algún consejo, y realmente lo aprecié. Jose y Almudena, gracias por el tiempo que pasamos juntos, especialmente el surf!

Mi tiempo en el CABD ha sido espectacular, principalmente gracias a gente como Rafa, Marta, Marta Peque, Ana, Jose Luis, Sandra, Jose Maria, Alvaro, Alejandro, Juan Tena, Joaquín y Laura, los gatitos y los yahtzee cada dia. Quiero agradecer especialmente a Ibai y Panos, mis mejores amigos y colegas bioinformáticos, por ser grandes frikis y ciclistas, por todas las interminable charlas al pescadito o en Casa Eme. Un abrazo a Carla, Txula y Argi, por hacerme sentir parte de vuestra familia y por todo vuestro cariño. ¡Nos encontraremos de nuevo en Berlín! Me gustaría agradecer también a Isabel y a Nacho, por ser tan buenos amigos, mentores, foodies, feministas, brillantes científicos y siempre solidarios durante estos años. ¡Les deseo lo mejor y llevaré esa receta de pollo coreano en mi corazón donde quiera que vaya! Sevilla ha sido mi hogar durante los últimos tres años, y algunos de sus vecinos me hicieron como en casa practicando español conmigo, o conversando mientras me servían cañas. Mamen, Leo, Eme, Riccardo, Marco, Miguel, gracias. Hablando de cañas, extrañaré nuestras Cañas con Caitlin, Gang. Esos martes por la noche de cubos, aceitunas y charlas con Ben, Madi, Alexandria, Kendall, Ofelia, Amanda y Francesca son algo que definitivamente echaré mucho de menos. Gracias por todos los memes, los videos increíbles, los picnics con vistas al rio. Gracias Sevilla por ser una ciudad loca, calurosa y maravillosa, llena de vida, arte y gilipollas en el carril bici. Lo que tienes un color especial no es broma. Te echaré mucho de menos.

Un ringraziamento particolare va a Matteo, Gianmarco, Lorenzo, Rino, Anna, Timothy e Laura. Ogni volta che torno a casa ci siete sempre e trovate una serata per incontrarci. Una mezz'ora, uno spritz da Tocc e sembra che non sia mai andato via. Grazie per esserci, soprattutto perché avete le vostre vite e riuscite sempre a ritagliare un po' del vostro tempo per me.

I'd like to also thank Cindy and Brent, Christy and Cole, for welcoming me in your family. I love you guys! Thanks for supporting me, for the mountains and the hikes, for the turkey at Christmas, the jokes and the wonderful time spent together. I'm lucky I got that hiking gear! I'm blessed to have all of you in my life.

Grazie al Nonno Giorgio, alla Nonna Flora, al Nonno Mario, per essere dei nonni meravigliosi, per le storie, i Natali insieme, il lavoro nei campi e il caffè alle macchinette a Valdobbiadene. Siete una roccia.

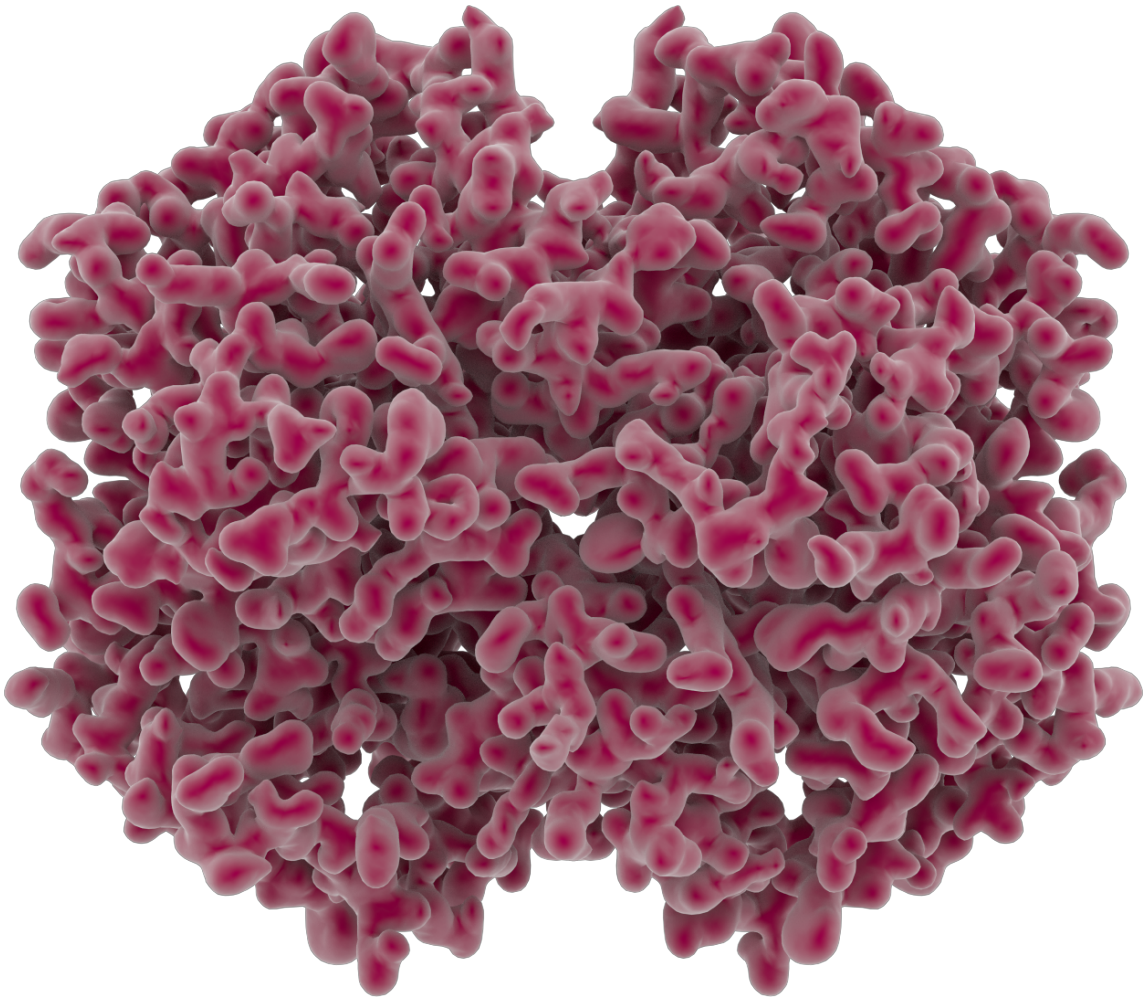
Uno struccone a Elisa e Fede, viaggiatori, artisti, amici e fratelli. Vi auguro tutto il bene e la fortuna possibile. Ci vediamo presto! Grazie anche ad Ale, per sopportare mia sorella! A parte gli scherzi, sono veramente felice di averti in famiglia.

Un abbraccio forte a Mamma e Papà, per avermi supportato e sopportato tutti questi anni. Non importa dove andrò dopo, ma so che ci sarete sempre al mio fianco a consigliarmi e a farmi da supporto morale. Vi voglio un mondo di bene.

This thesis is dedicated to us, Cait. You turned my life into something amazing, packed with travels and love, pictures and chats, books and music. You're an incredible human being and I consider myself so lucky to have you as my partner. Thank you so much for helping me through the thesis, the proofreading, the helpful comments. You make me a better writer, scientist and person. You make me smile and laugh, and you're my strongest supporter. Thanks for believing in me. I love you.

Nicola

Table of Contents



Resumen/Summary

Resumen

Las proteínas son la clave para entender la biología celular. La determinación de su rol y función nos ayuda a descubrir las características de los procesos moleculares en la base de la vida. Las técnicas de alto rendimiento han permitido a los científicos acumular una gran cantidad de datos sobre secuencias de ADN de miles de organismos diferentes. La función de las proteínas codificadas en estas porciones de ADN se determina por métodos de anotación manuales o automáticos, utilizando experimentos computacionales y biológicos para obtener una descripción coherente. Aunque la revisión manual de estas predicciones finalmente produce las anotaciones más fiables, este enfoque no es factible con la tasa actual de secuencias depositadas en las bases de datos biológicas. Esto afecta el conocimiento de la biología de varios organismos.

Los esfuerzos de revisión manual se centran principalmente en la caracterización de organismos modelo En consecuencia, las bases de datos donde se reúne la información abarcan grandes cantidades de datos para un subconjunto específico de organismos. Actualmente, solo los grandes consorcios pueden generar estos recursos web, mientras que otros grupos que investigan organismos recientemente secuenciados carecen de los medios y recursos para lograr una anotación de proteoma más completa. Además, la gran mayoría del software para anotación de proteínas se enfoca solo en algunos aspectos de la función de una proteína; por lo tanto, falta información complementaria que podría derivarse de otras fuentes, tanto in silico como in vivo.

El objetivo de esta tesis es desarrollar un nuevo enfoque para la anotación de funciones de proteínas que aborde los problemas mencionados anteriormente, incluidas nuevas herramientas y recursos para mejorar el estado actual en el ámbito de la predicción de la función, para así aplicarlo a organismos no modelos. Lo llamamos “Integrative Cell Biology” (ICB) o Biología Celular Integrativa.

ICB se basa en la integración de varias fuentes de datos, incluyendo características de secuencia y estructura. De esta forma podemos obtener una anotación más amplia que proporciona al usuario una descripción más completa de una proteína. ICB

también es capaz de visualizar múltiples proteínas de una manera fácil y rápida a través de un navegador web.

Probamos el enfoque Integrative Cell Biology con una “pipeline” computacional resultante para caracterizar 39 proteomas del superfilo bacteriano *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC). Además de su relevancia en varios campos, sus proteomas tienen un bajo porcentaje de proteínas anotadas, y solo unas pocas se han caracterizado experimentalmente. Sus propiedades fueron determinadas por observaciones experimentales, mientras que las secuencias que las codifican son en su mayoría desconocidas.

Al aplicar el pipeline ICB, aumentamos drásticamente la cantidad de anotaciones de sus proteomas, abordando cuestiones biológicas sobre su comportamiento.

Con el fin de hacer que nuestros hallazgos estén disponibles para la comunidad de investigación de PVC, creamos PVCbase, una plataforma única para examinar los resultados de ICB a través de DataTables, realizar búsquedas de secuencia basadas en homología y visualizar las características de la estructura secundaria de las proteínas.

Para demostrar aún más las capacidades de ICB, analizamos tres Planctomycetos recientemente secuenciados asociados al entorno de macroalgas. *Los* genomas de *Rubripirellula obstinata* LF1, *Roseimaritima ulvae* UC8 y *Mariniblastus fucicola* FC18 se ensamblaron, se anotaron utilizando ICB, y se caracterizaron adicionalmente comparándolo con Planctomycetes de otros ambientes. Posteriormente se complementaron sus rutas metabólicas y se evaluó su identidad a través de la filogenia. Tras los análisis pudo verse que algunas proteínas están involucradas en la interacción con los hospedadores de algas, incluidas algunas de tamaño extraordinario que merecen un análisis posterior.

Se creó una versión de contenedor Docker de ICB que agiliza la instalación y el uso de pipelines, permitiendo que los grupos de investigación con intereses compartidos creen una plataforma similar a PVCbase. La salida de DataTables y la diversidad de herramientas incluidas permiten una transición fluida de secuencias a anotaciones de proteínas fácilmente navegables. Estos recursos crean entornos compartidos para analizar grandes conjuntos de proteínas, con poco o ningún conocimiento de codificación requerido.

El concepto de Biología Celular Integrativa y sus recursos derivados contribuyen al campo de la predicción de la función de la proteína y proporcionan una solución en el caso de organismos mal anotados o recién secuenciados. PVCbase ha sido utilizado por varios grupos de investigación en microbiología de PVC (16 universidades de 14 países hasta agosto de 2018) y su base de usuarios se beneficiará de la adición de proteomas y de los análisis. Integrar varias fuentes de información para evaluar la función de la proteína es una posible solución a la inconsistencia y falta de fiabilidad de las herramientas de predicción. Al utilizar ICB, podemos responder preguntas que no podrían abordarse por otros medios. En el futuro, nuevas fuentes de información implementadas en ICB ampliarán nuestro conocimiento de varias características desconocidas de varios organismos.

Summary

Proteins are the key to understanding cell biology. Determining their role and function helps us to discover the features of molecular processes at the base of life. High-throughput techniques have allowed scientists to amass a vast amount of data on DNA sequences from thousands of different organisms. The function of proteins encoded in these portions of DNA is determined by either manual or automated annotation methods, using computational and biological experiments to obtain a coherent description. Although manual curation of these predictions produces the most confident annotations, this approach isn't feasible with the current rate of sequences deposited in biological databases.

Manual curation efforts are mostly focused on characterizing model organisms, resulting in centralized hubs that encompass vast amounts of data for a specific subset of organisms. Only large consortiums are able to generate these web resources, while other groups researching newly sequenced organisms lack the means and resources to achieve a complete proteome annotation. Furthermore, the vast majority of software for protein annotation purposes are focused only on a few aspects of a protein's function; therefore missing some complementary information that could be derived from other sources, both in silico and in vivo.

The aim of this thesis is to develop a new approach to protein function annotation that addresses the issues mentioned above, including new tools and resources to improve the current status of the field to apply to non-model organisms.

This approach relies on the integration of several data sources, including sequence and structure features, to obtain a broader annotation that provides the user with a more complete overview of a protein. It also visualizes multiple proteins in a easy and rapid manner through a web browser. We called it Integrative Cell Biology (ICB).

We tested the Integrative Cell Biology approach with a resulting computational pipeline to characterize 39 proteomes from the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) bacterial superphylum. Besides their relevance in several fields, their proteomes have a low percentage of annotated proteins, with only a few being

characterized experimentally. Most of their interesting features are encoded in sequences that are unknown or only partially.

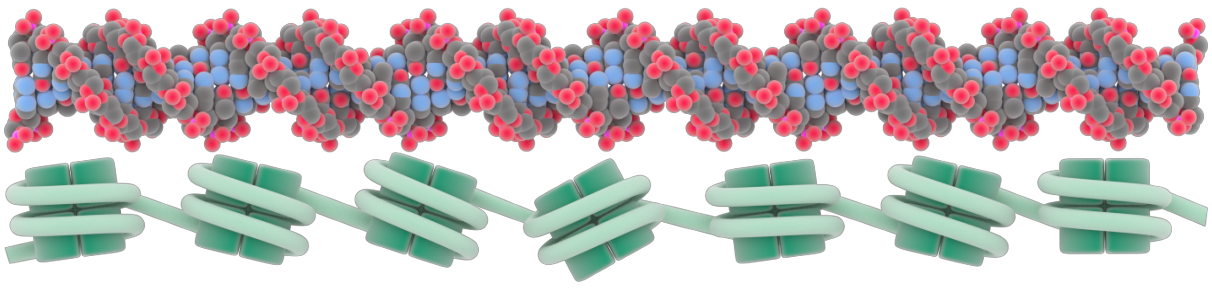
By applying the ICB pipeline, we drastically increased the amount of annotation of their proteomes, addressing biological questions on their behaviour. We then developed tools to characterize them further. In order to make our findings available for the PVC research community we created PVCbase, a one-stop platform to browse the results from ICB through DataTables, perform homology-based sequence searches, and visualize proteins' secondary structure features.

To further demonstrate ICB's capabilities, we analysed three newly sequenced Planctomycetes associated to the macroalgal environment. *Rubripirellula obstinata* LF1, *Roseimaritima ulvae* UC8, and *Mariniblastus fucicola* FC18 genomes were assembled, annotated using the ICB pipeline, and furtherly characterized by comparing them with Planctomycetes from other environments. We then complemented their metabolic pathways and assessed their identity through phylogenetics. We found that some proteins are involved in the interaction with algal hosts, including some with extraordinary size that deserve further analysis.

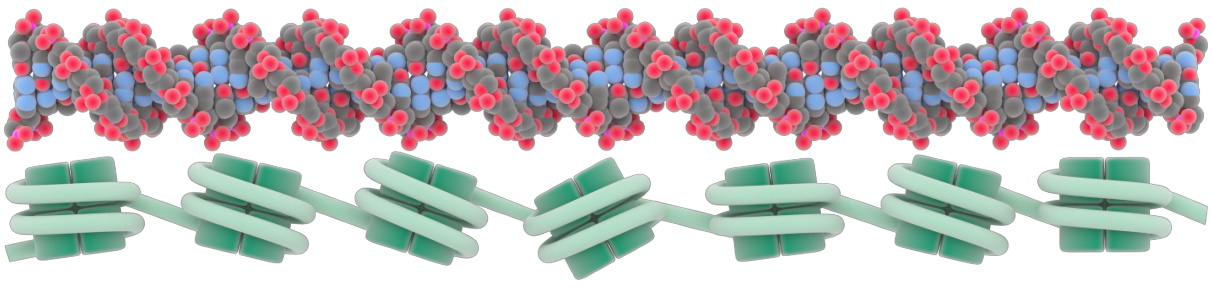
We created a Docker container version of ICB that streamlines pipeline installation and usage, allowing research groups with shared interests to create a platform similar to PVCbase. The DataTables output and the diversity of tools included allow a smooth transition from sequences to easily browsable protein annotations. These resources create shared environments for analyzing large sets of proteins, with little to no coding knowledge required.

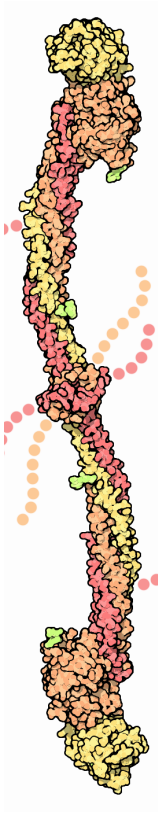
The Integrative Cell Biology pipeline and its derived resources contribute to the field of protein function prediction and provide a solution when dealing with poorly annotated or newly sequenced organisms. PVCbase has been used by several research groups in PVC microbiology (16 universities from 14 countries as of August 2018) and its user base will benefit from the further addition of proteomes and analyses. Integrating several sources of information for assessing protein function is a potential solution to the inconsistency and unreliability of protein function prediction tools.

Using ICB we can answer questions that couldn’t be addressed by other means. Going forward, newer sources of information implemented in ICB will further our knowledge of several unknown features of various organisms.



Introduction





1 Molecules

1.1 Proteins: the building blocks of Life

Life is based on chemistry. Molecules such as nucleotides, amino acids, lipids and polysaccharides are the foundation of every living organism on Earth. These building blocks, when combined in a linear fashion as DNA, RNA, and proteins, create the fundamental structures needed by a cell for its function, reproduction and ultimately heredity, over time. In particular, proteins are involved as core components and actors in almost every cell process. From DNA replication to carrying out important enzymatic reactions, importing nutrients to the cell, and transmitting signals, these macromolecules are what defines a specific pathway or process inside a cell. In bacteria, the presence or absence of an enzyme (a protein that transforms a substrate into other molecules) can influence their survival in a particular environment. The flexibility of proteins embedded in a cell membrane allows the entrance of specific metabolites, while also regulating the cell's resistance to heat or chemical shocks (**Figure 1A**).

Through lateral gene transfer, bacteria are able to inherit genes and therefore proteins, resulting in new functional features. Antibiotic resistance, surviving on a different substrate, and interacting with a host is regulated by specific enzymes. Different sets of proteins in an organism's proteome can sway the behaviour of an organism, from symbiosis to pathogenicity.

Proteins on the outer surface of a cell create an interface with other organisms or within the same species, resulting in the creation of biofilms or an immune response (**Figure 1B**). In the case of viruses, we could say that their entire existence relies mostly on proteins. A bacteriophage iconic "head" is made of a protective protein coating that contains its DNA and RNA. The entirety of the phage body, including the collar, sheath, and injecting needle are made of proteins too (**Figure 1C**).

Viruses like the flu rely on two proteins, haemagglutinin and neuraminidase, to recognize its host and use them to lyse the cell for infection. A change in their structure, due to sequence changes, can alter their affinity for a specific host, resulting in species

jumps that can be extremely dangerous. With a few amino acids substitutions, the avian flu can infect humans, and with less than 10 substitutions and insertions, there would be another Spanish Flu-like strain epidemic (Imai et al, 2012).

DNA takes most of the limelight on the news, but the proteins that are encoded in it are the underlying motor of the cell and life as we know it.

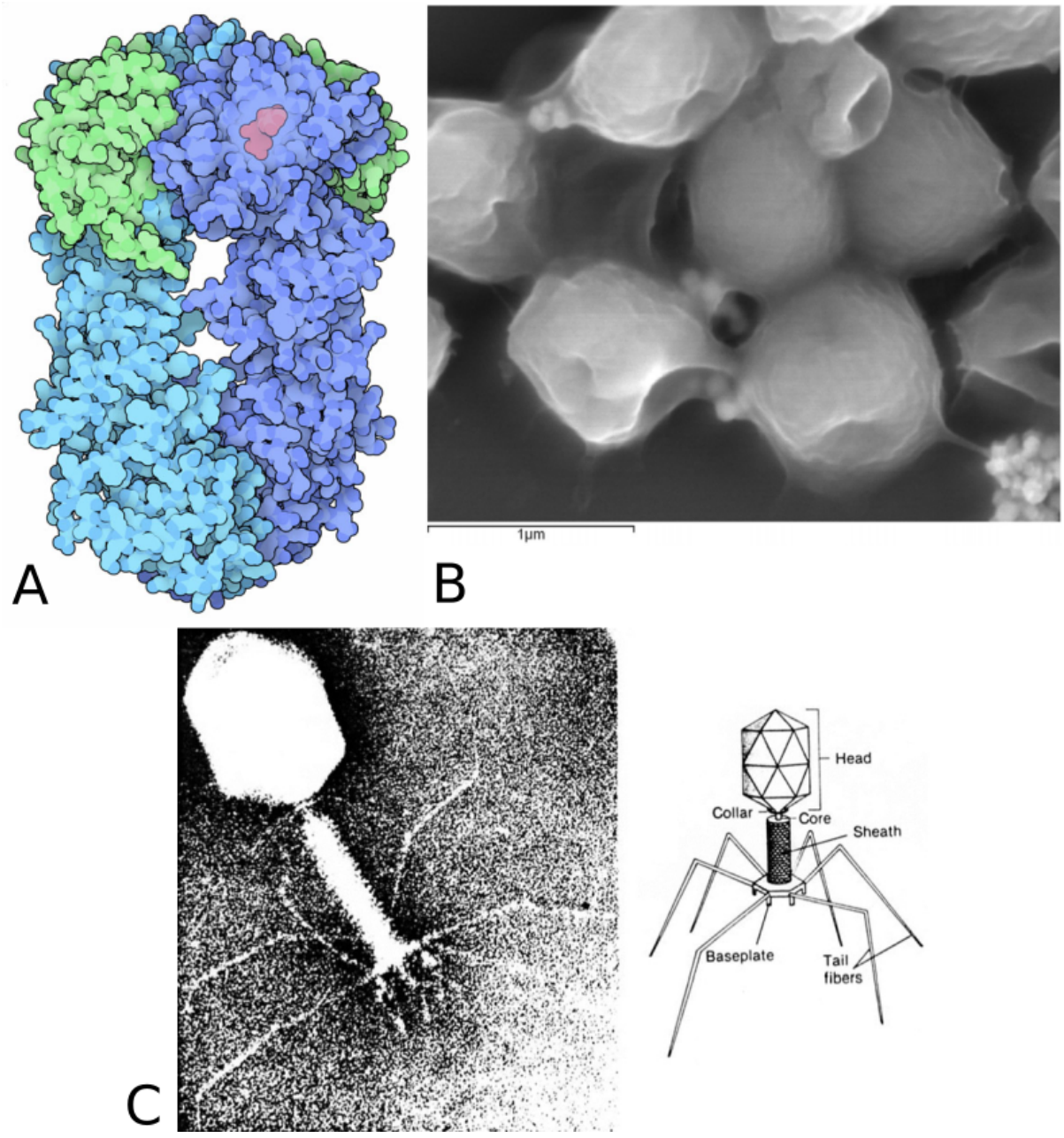


Figure 1: A. Hsp90 (blue) and cochaperone Sba1 (green) with bound ATP (red) (Source:PDB101). **B.** Bacterial biofilm on the surface of macroalgae (Faria et al., 2017). **C.** Electron Micrograph of bacteriophage T4. Right. Model of phage T4.

1.2 Genomes and Next Generation Sequencing

Although DNA's structure was reported in 1953 (Watson and Crick, 1953; Franklin and Gosling, 1953; Wilkins, Stokes and Wilson, 1953), it wasn't until 40 years later that the first complete genome of a free-living bacteria was sequenced. In 1995 the complete genome (the sequence of all DNA contained in a cell) of *Haemophilus influenzae* Rd was released to the public community. With a size of 1,830,137 base pairs arranged in a circular chromosome and 1743 protein-encoding genes, it was the very first organism's genome to be completely sequenced (Fleischmann et al., 1995). Since then, more than 40 thousand complete genomes have been made available to the public to uncover the foundation of these organisms and the number is increasing by the thousands every year. This trend is ongoing since the introduction of shotgun sequencing and, more recently, several techniques dubbed either Next Generation Sequencing (NGS) or High-Throughput Sequencing. These techniques use different methods, from pyrosequencing (Roche 454) to the newest nanopore sequencing like the Oxford Nanopore. What they have in common is the generation of fast and redundant reads that provide fewer sequencing errors and good "coverage" of an organism's genome. Competition and the ability to sequence genomes on a massive scale have caused a dramatic drop in sequencing costs. The pioneering Human Genome Project, created to produce the first draft of the human genome, had a final price tag estimated at around 2.7 billion dollars. The same feat can now be obtained for a little more than a 1000 US Dollars. Since the introduction of NGS, due to the lower price and the higher throughput, the number of sequences and base-pairs that have been deposited in Genbank and Whole Genome Shotgun Projects (WGS) skyrocketed (**Figure 2**).

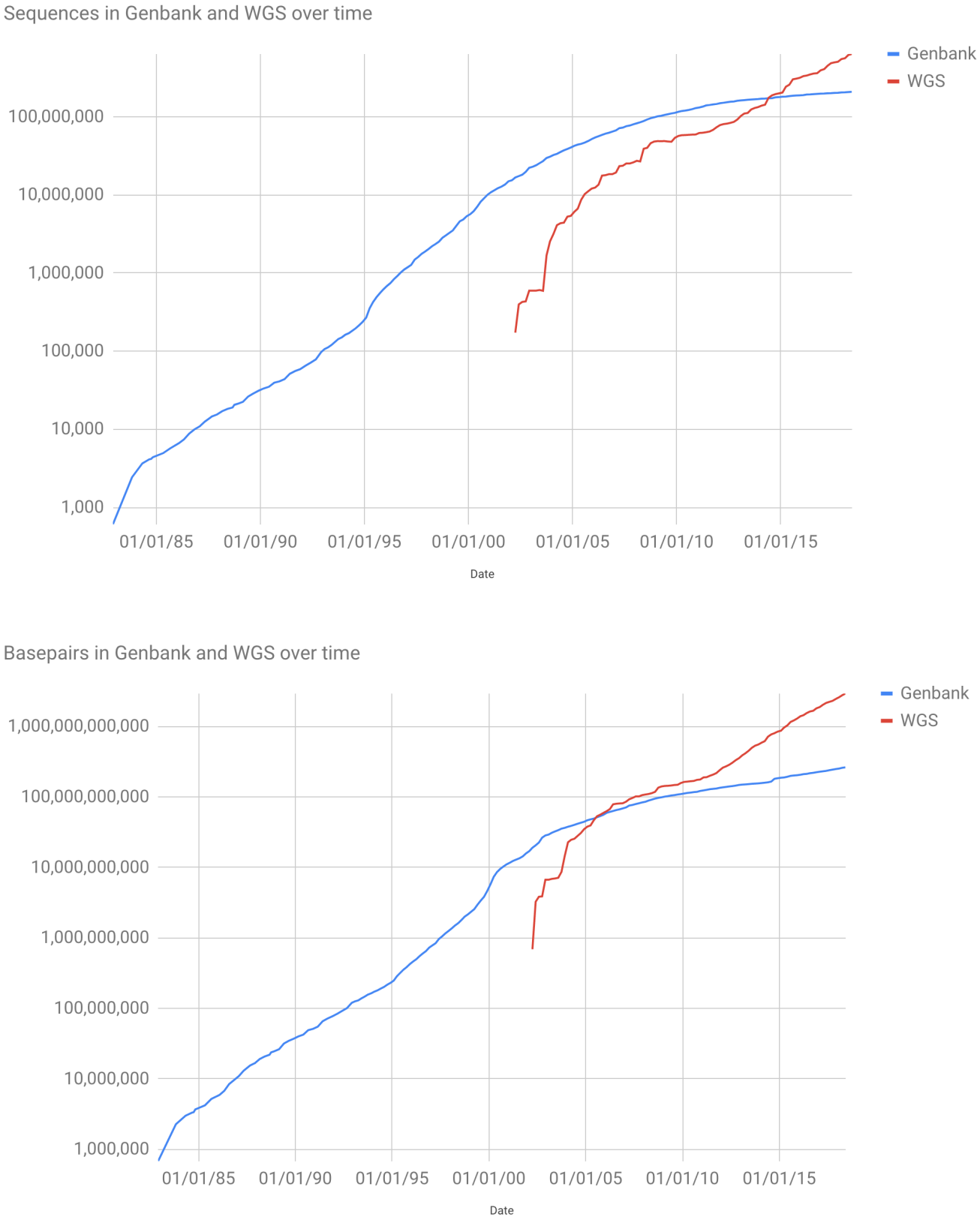


Figure 2: Number of deposited sequences and base-pairs in Genbank and WGS. The top graph shows the increase of single sequences (genes or transcripts) deposited in Genbank and WGS. The bottom graph shows the number of total base-pairs that have been deposited in Genbank and WGS.

However, obtaining the genome sequence is just the first of many steps towards understanding the workings of the cell. Once we obtain the raw sequencing reads, the genome needs to be assembled and its genes annotated.

Gene annotation is challenging in general, but even more in the case of eukaryotic genomes, where introns are involved, and the coding portion is sparse throughout the chromosomes. This issue is not present in the case of Bacteria and Archaea, due to the complete lack of introns and the presence of genes in well-defined regions called operons.

Protein-encoding regions are discovered using several sequence features, such as organism-specific codon usage, poly-A repetitions, and start codons. Specific regions in their vicinity, like Shine-Dalgarno ribosome binding sites and promoters, give hints to the proximity of a gene but also about their regulation. These approaches, although correct in most cases, tend to have some signal noise and overpredict the number of Open Reading Frames (ORFs).

In order to solve these issues, information from the aforementioned approaches are combined with homology-based methods in order to obtain a more realistic gene prediction.

Recent annotation tools, like Prokka (Seemann, 2014) for prokaryotes and Augustus (Keller et al., 2011) for eukaryotes, use a combination of empirical methods based on homology, ab-initio predictions that include probabilistic methods like Hidden Markov Models (HMM), and artificial neural networks.

1.3 Proteins sequence and structure

When a gene is expressed, a complex machinery called the RNA polymerase II, transcribes DNA into RNA which eventually gets translated into a protein by a ribosome.

Starting with a Methionine codon, every triplet of nucleotides in RNA are associated with either a specific amino acid or a stop codon that signals the ribosome that the translation of that specific gene is complete (**Table 1**). Each amino acid can be grouped into five major categories, based on the chemical properties of its side chain. These are nonpolar-aliphatic, polar-uncharged, aromatic, positively charged, and negatively charged.

	U	C	A	G	
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
	UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GGA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Nonpolar, aliphatic Polar, uncharged Aromatic Positively charged

Negatively charged

Table 1. The genetic code. Each RNA triplet corresponds to a specific amino acid, or a signal that terminates the gene translation.

These characteristics are fundamental to the protein function and localization. Amino acids that have a sulfur atom in their lateral chain, like cysteine and methionine, are able to create a disulfide bond that stabilizes two portions of the protein. In order to be inserted in a membrane, amino acids with a high degree of hydrophobicity are required in its spanning region, while charged or polar ones are present in the inner core of ion channels and metabolite pumps. Specific transporters are able to import different substances based on the amino-acids affinity for the transported molecule. Modifications through kinases and phosphatases regulate the access for these substances through a conformational change in the protein. The protein's three higher orders of structure are defined by its amino acid sequence (the primary structure) (**Figure 3**).

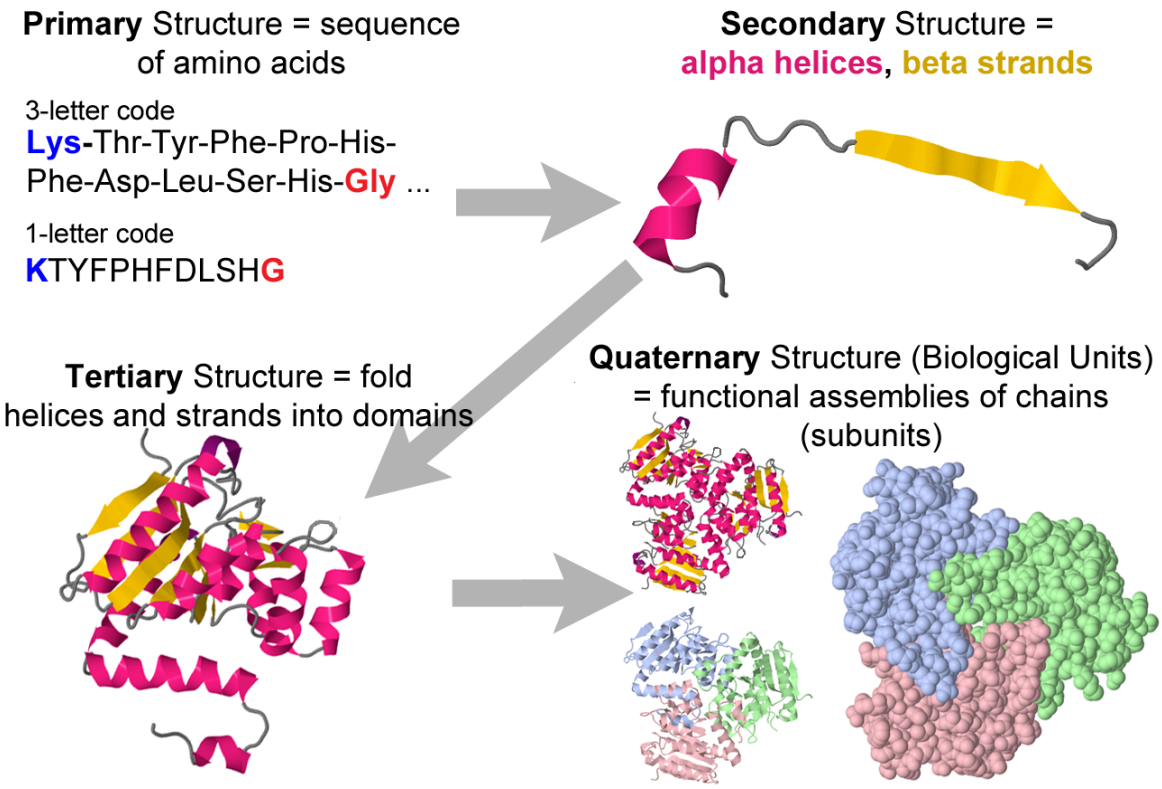


Figure 3. Protein structure levels (Source Proteopedia. http://proteopedia.org)

The primary structure consists of the linear sequence in which the amino acids are connected through covalent bonds. The secondary structure is determined by the creation of hydrogen bonds between neighboring residues, resulting in the formation of loops, helices, and strands connected by non-structured regions.

These particular arrangements form the tertiary structure, where helices and strands are grouped in domains that minimize the protein's free energy and stabilizes the protein overall. Each of these domains does not require further stabilization and are independent of each other; their combination in multimers of several chains generates the quaternary structure.

1.4 The protein knowledge gap

DNA and protein sequence databases have experienced exponential growth during the past decade (O'Leary et al, 2016). A downside of this phenomenon is that the rate of experimental characterization cannot match the ever-growing amount of biological sequences (Erdin et al, 2011) **(Figure 4)**.

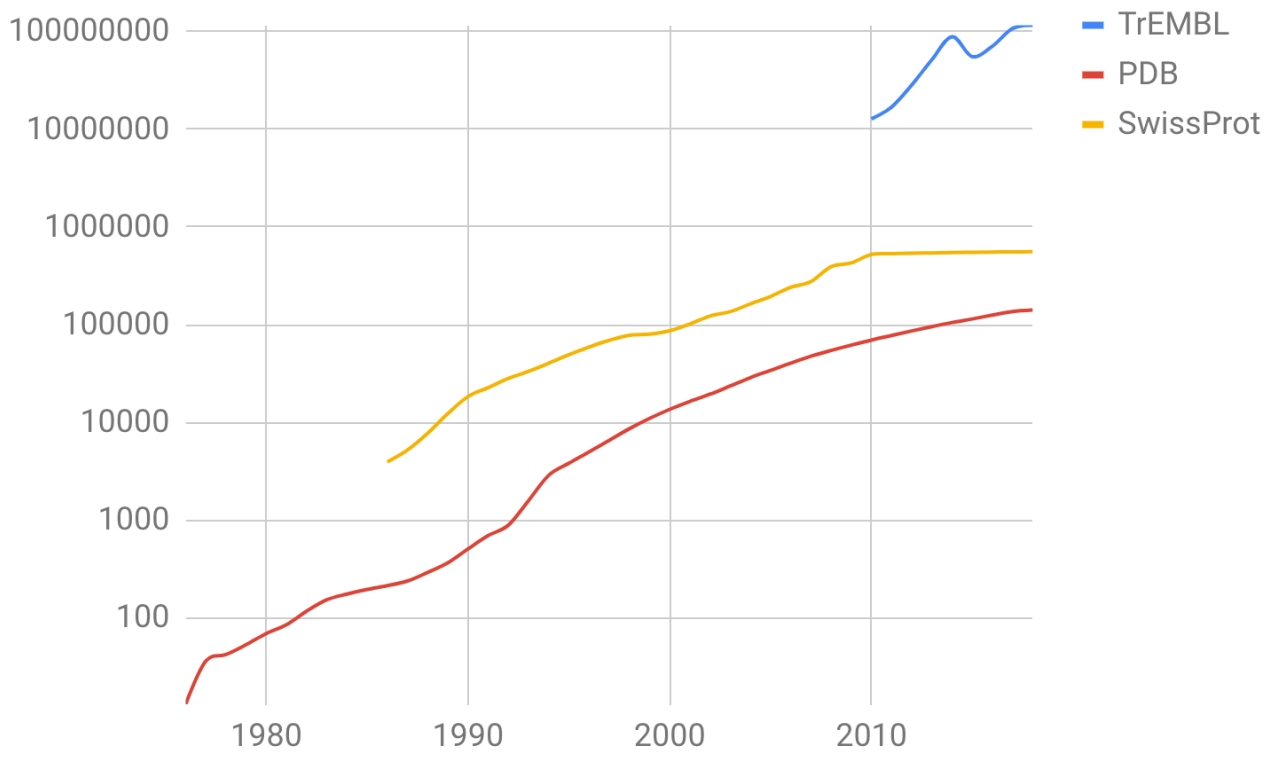


Figure 4. Growth of protein databases over time.
In blue, sequences deposited on the TrEMBL (automatic annotation).
In yellow, sequences manually annotated in SwissProt.
In red, protein structures deposited in PDB.

Therefore, there is a pressing need to translate the sequence data into comprehensible information and create tools to help in the effort.

One of the first problems that arise in determining protein function, is the definition of protein function itself.

A protein's function is defined by its structure, including specific modifications of its amino acid sequence, the timing and location of its expression, and other factors (interactions, activation, regulation, etc.). Some proteins are involved in maintaining the cell shape, others can modify their conformation to react to stimuli or transfer information to signal to the cell a change in its environment. While these characteristics are intrinsic to the protein itself, its overall function is related to the general cell environment. When and where a protein is expressed, the interactions with other protein complexes or processes, and the protein state are what define a protein's function. Therefore, a protein's role can change depending on multiple factors.

Many proteins with similar sequences and structure usually have a similar function, but there are exceptions. Some proteins are similar but act completely different from each other (Gerlt and Babbitt, 2001). On the other hand, there are opposite examples of non-homologs with convergent evolution, resulting in proteins with similar function and different structure (Galperin and Koonin 2012, Omelchenko et al., 2010). Due to shared similarities, a protein's function can be inferred from other characterized proteins using bioinformatics tools.

Due to the volume of data, most protein functional annotations are performed automatically by computational methods. However, these cannot assign a function to all proteins being analyzed, always leaving a fraction of the proteins lacking any significant functional information (*unknown function, uncharacterized* and *putative* proteins represented 38% of TrEMBL as of September 2018). Most computational methods of functional assignment are based on homology, assuming that proteins derived from a common ancestor will be functionally related at some level (Loewenstein et al, 2009). However, their function may have deviated considerably since duplication and divergence from their common ancestor. Thus, the amount of functional description that can be transferred from one annotated protein to an uncharacterized one is variable and related to the evolutionary relationship between the two proteins. The evolutionary relationship between two given proteins can be determined by sequence comparison. However, depending on the specificity of the functional description, sequence similarity may not always be sufficient to justify transferring function from one protein to another (Devos and Valencia, 2000). Proteins can also have more than one function (Huberts, Van der Klei. 2010), due to gene

fusion, which further complicates the task of assigning a function to these polypeptides. One consequence of this process is the propagation of incorrect annotations when a functional assignment is not detected or when an erroneous assignment is transferred to another protein. Because of this, databases inevitably contain errors in their annotations which are extended to new proteins by automatic computational analyses (Schnoes et al., 2009).
Still, automated large-scale annotation exercises by homology represent the first level of functional assignment. Most of the time it is also the only functional assignment given. Despite the obvious limitations tied to function predictions based on sequence features only, this method has also been incredibly successful and has contributed to many biological discoveries in the latest decades.

1.5 Current status of Protein Function Prediction

Function prediction has improved significantly during the last decade. From the first attempts based only on BLAST sequence similarity, homology-based tools improved by adding layers of information to these alignments. Recent versions of BLAST, like PSIBLAST (Altschul et al., 1997) use sequence profiles based on features present in the protein, improving the detection of distant relatives. Besides BLAST, newer approaches like HMMER (S.R.Eddy, 1998) and HHblits (Remmert et al., 2011) compare Hidden Markov Models generated for each sequence, with additional information coming from structure and domains. Tools shifted from sequence to pattern and motifs searches (PROSITE), domain localization, and conservation (ScanProsite) to multiple annotations from different sources such as InterProScan. Another improvement was introduced with the consideration of 3D structure similarity, which has more biological significance than sequence-based analysis provided by previous tools.
The increase in available resources and understanding of the subject is welcome, but it has a downside. Each one of these tools has a narrow field of application, and the information provided can be incomplete, uninformative, and redundant. Many predictors often try to ascribe the full function of a protein with a single approach without considering the multiple aspects of protein biology.

Many current computational function prediction methods aim to define the most specific function and reach a conclusive prediction. Specific databases contain the results of these single resources (Superfamily, ELM). However, these results are too focused and efforts in merging them are scarce. The RefSeq (O'Leary et al., 2016, Tatusova et al., 2014) and UniProt (The UniProt Consortium 2015) resources collect different sources of protein characterization for each protein, but most results are obtained through automatic annotation and only a fraction of entries are manually-curated.

Sometimes a consensus between different analyses can be reached but working at this specific level can make the task cumbersome and time-consuming; having a more general overview from the start makes the task easier and faster. A better annotation environment can be created by addressing some issues that are currently hindering the field.

A major issue is the lack of automatic communication and translation between tools and outputs. In genome bioinformatics, new suites are emerging with built-in interconnectivity, as well as standard outputs that can be imported in subsequent analyses. A common framework allows results from different predictors to be implemented with effective cross-talk between platforms with different aims.

Contrary to DNA, information on the protein annotation side does not only rely on sequence and structure-based methods, making the task of automating predictors and finding a standardized language difficult. Future annotation tools will have to rely on the integration of different methods involving many protein features like sequence, domains, structure, interactions, subcellular location, and metabolism (Earnshaw, 2013). Some existing efforts include ANNIE (Ooi et al., 2009), BAR PLUS (Piovesan et al., 2011), and others (Tiwari and Srivastava, 2014; Yamada et al., 2012).

Other issues with the current status of protein annotation tools and databases are accessing these resources and comprehensively visualizing the results.

Most databases require multiple queries or comparing different files, restricting their utility for "wet lab" scientists or large-scale analyses. Integration and efforts in data visualization are necessary for the future of the field and should be encouraged.

1.6 Integrative Cell Biology

Our approach to solve the issues mentioned above, and the scope of this work, is called Integrative Cell Biology, or ICB. ICB is a computational pipeline for characterizing large numbers of proteins with an increased accuracy compared to previous methods that rely on one or limited tools. In order to correctly pinpoint the role of the protein in the cell, we need to integrate different sources of information that describe multiple aspects of the protein from in-silico and in-vivo predictions. The resulting Gene Ontologies (GO) describe each protein's role in the cell´s processes, where it's expressed, and the activity it carries out. The pipeline output addresses some of the issues with data visualization by allowing quick and easy browsing. The user can annotate several proteins at once, making it suitable for the characterization of newly sequenced organisms.

Information that other pipelines often overlook is collected and used to obtain a more complete and confident function prediction.

Examples of information taken into account are that the presence of signal peptides give us information about the protein's localization outside of the cell, while if a specific number of transmembrane helices are encoded in the sequence it gives clues that the protein's function is probably related to membranes and trafficking. Disorder prediction provides the user with valuable knowledge on the globularity of the protein and an indication if the protein is flexible, related to cell trafficking or its amenability to crystallization (Pietrosemoli et al., 2013, Tantos et al., 2012, Busch et al., 2015).

Using a variety of predictors and integrating the results, with ICB we can attempt the identification of proteins that have been defined as "uncharacterized" or "protein of unknown function".

Our function prediction of the protein WP_010034877.1 from the bacteria *Gemmata obscuriglobus* is an example of successful application of the ICB approach.

Annotated as a hypothetical protein in GenBank, this particular protein doesn't show any result using the PSIBLAST module but using a combination of the other modules provided information to assess its function. The HHblits module describes the protein as an Alpha-ribazole phosphatase and this prediction is supported by the output of HHpred, telling us that the structure is almost identical to a Phosphoserine phosphatase 1-hydrolase, a closely-related protein. InterProScan maps various

domains that are coherent with a phosphatase activity, such as being a member of the Histidine phosphatase and Phosphoglycerate mutase-like superfamily. Additionally, the pipeline predicts a lack of signal peptide and transmembrane helices, along with complete globularity, in agreement with the protein function and its predicted 3D structure (PDB code: 1HSK). (**Figure 5**) (**Figure 6**).

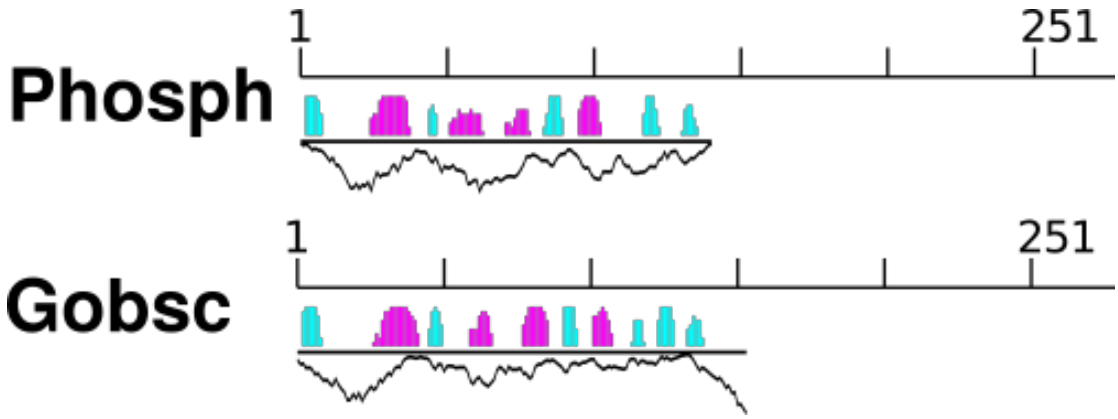


Figure 5. Predicted secondary structure of WP_010034877.1 (Gobsc) and a member of the phosphatase family (D3DFG8). The predicted α-helices (magenta) and β-sheets (cyan) are indicated by bars above each line. The height of bars is proportional to the confidence of the prediction. The zig-zagging line underneath depicts the disorder level for each amino acid.

Figure 6. WP_010034877.1 3D model obtained through HHpred (Remmert et al., 2011) and Modeller (Sali and Blundell, 1993).

Gemmata obscuriglobus	WP_010034877.1	
MODULE	RESULTS	
Initial Parser: length	154aa	
SignalP	0 [SP=n, topology=o]	
IUPRED	11.04 % [GlobDoms=(1-151)]	
HHblits	Alpha-ribazole phosphatase [Cov=86.4% Prob=100.0 Evalue=2.5E-36]	
HHpred	4ij5_A (1-146)	Phosphoserine phosphatase 1; hydrolase; [Prob=100.0 Evalue=3.2E-33]
InterProScan	PF00300: Histidine phosphatase superfamily (branch 1)	Histidine phosphatase superfamily (branch 1)
	G3DSA	3.40.50.1240
	SSF53254	Phosphoglycerate mutase-like
	SM00855	Phosphoglycerate mutase family

Table 2. Application of ICB pipeline for WP_010034877.1 and results for each of the pipeline modules

1.7 Disadvantages of Computational Pipelines

Pipeline

1 A line of pipe with pumps, valves, and control devices for conveying liquids, gases, or finely divided solids.

(Source: Merriam-Webster Dictionary)

2 In computing, a pipeline, also known as a data pipeline, is a set of data processing elements connected in series, where the output of one element is the input of the next one.

(Source: Wikipedia)

Like regular pipelines are a medium to transport something from a source to a destination through a pipe, computational biology pipelines are a collection of software blocks that transport and convert raw biological information from one tool to another until delivered as a result. While the foundation of modern NGS data analysis and almost omnipresent in bioinformatics research groups, computational pipelines for bioinformatics have some downsides. Most setups are difficult to manage, implement, and distribute, and require the installation of various modules, coding dependencies, and databases. This involuntary deep integration with the operating system discourages upgrading these pipelines to newer versions, resulting in deprecated predictions that affects annotation results.

Here are some current, practical disadvantages of computational pipelines for protein annotation:

- Most online databases are updated regularly, but this often doesn’t occur for local installations.
- Despite efforts in streamlining processes, most pipelines aren’t fully automatic, requiring manual intervention such as converting formats, launching tools, and parsing results.

- Most tools aren't optimized for laptops or desktop setups. On the other end of the spectrum, some programs aren't optimized for parallel processing or HPC (High Performance Computing) queueing systems.
- Most tools for DNA and protein annotation are built for UNIX environments like Linux and require knowledge of the command line. Efforts in development for other platforms, like Microsoft Windows and MacOS, are limited.
- The vast majority of wet lab biologists have limited knowledge of UNIX environments and their favorite working platforms are Windows and MacOS.
- Most experimental laboratories lack access to HPC infrastructures, forcing them to rely on external resources.

1.8 Software containerization

A potential solution to the problems presented above is a new technology called software containerization. Similar to shipping containers used to move goods, software containers are a way to reliably run tools when moved from one environment to the next. The container is a smaller version of a runtime environment, with its dependencies, libraries, scripts, and databases already pre-set, that runs without being influenced by the system infrastructure. Compared to virtualization tools like VMware or VirtualBox that requires multiple copies of the OS, one for each image, containers are built on top of an underlying shared OS, making them incredibly lightweight to run, store, and manage.

These advantages have prompted most of the current players in the tech industry (Google, Microsoft, and IBM, among others) to invest in the Open Container Initiative, releasing open source container systems to be used for server virtualization and creating multiple services to control them.

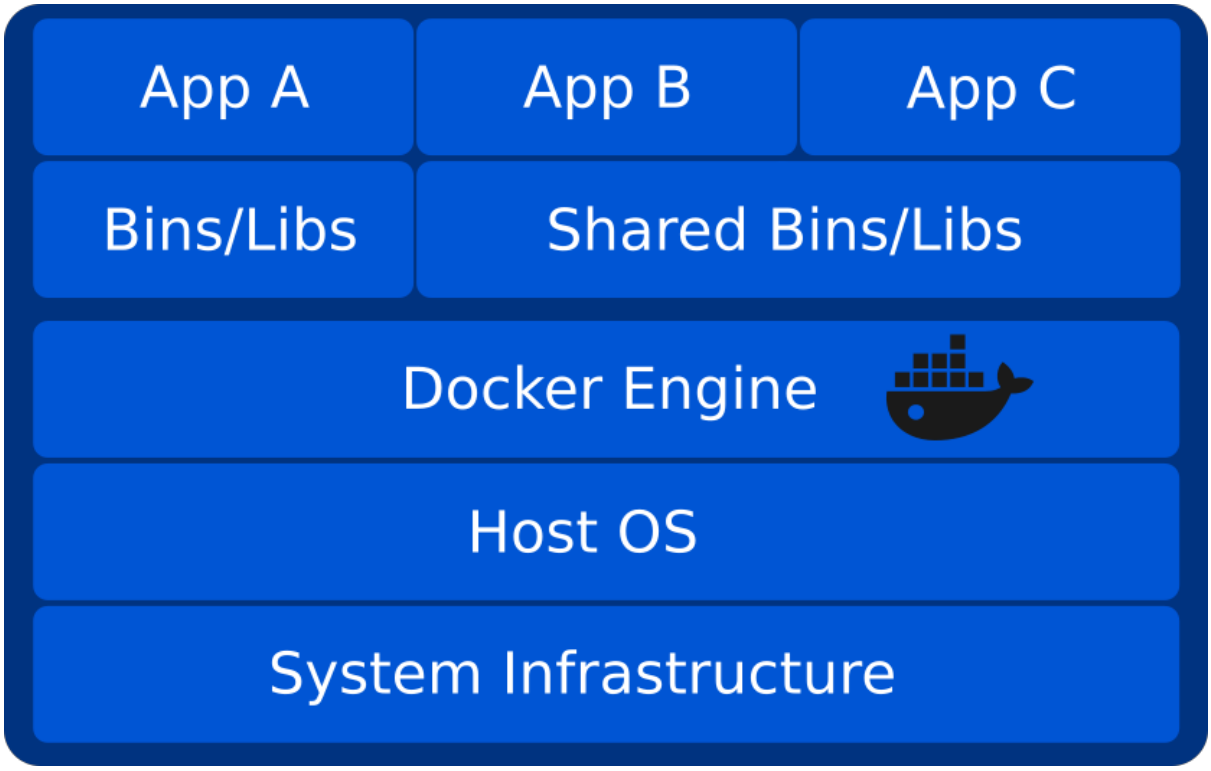
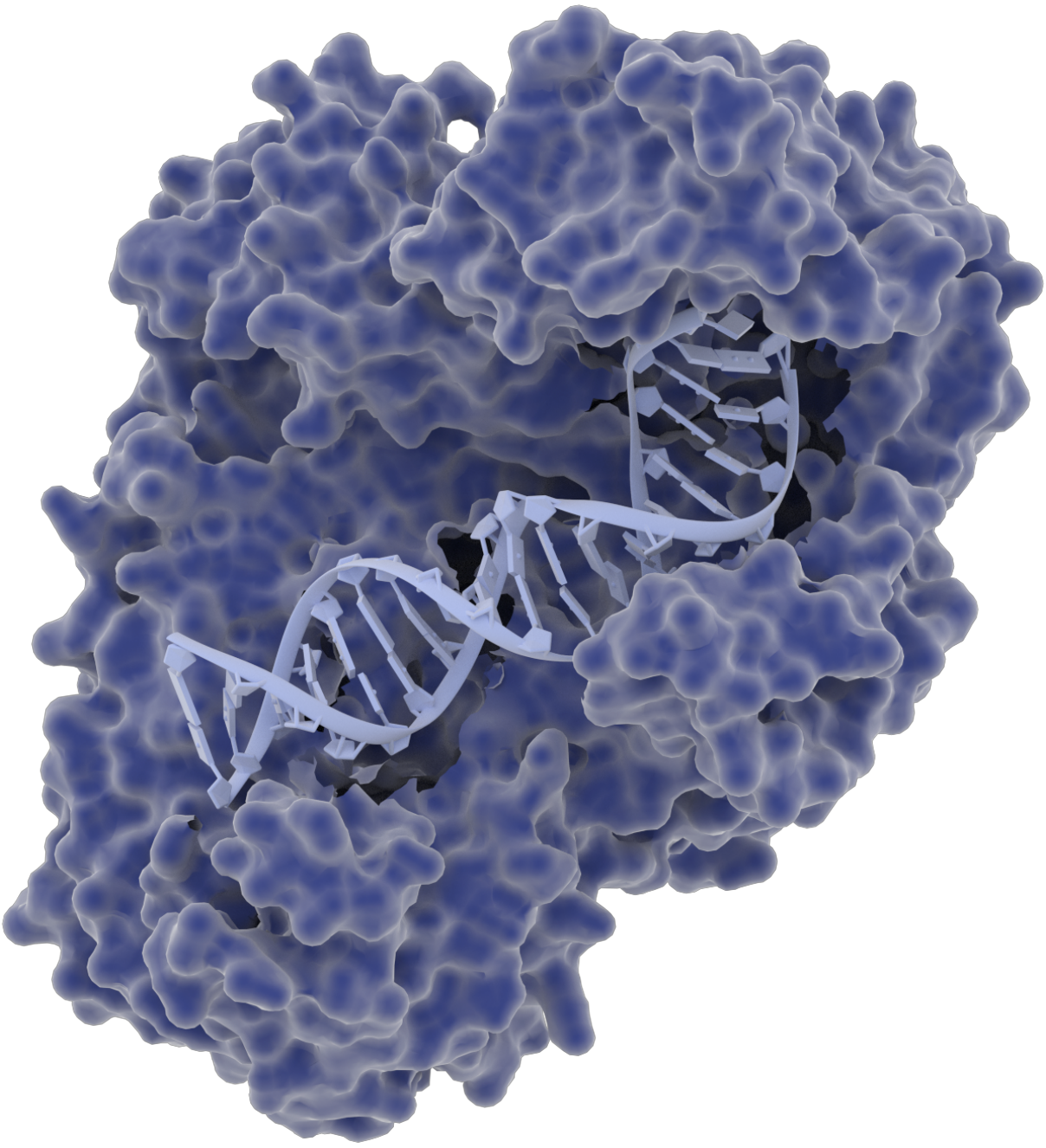


Figure 7. Docker system architecture.

On a smaller scale, container technologies like Docker (Merkel, 2014) allow cross-platform compatibility for various projects, allowing developers to write an application once without having to worry about inter-OS discrepancies. In the field of scientific software, it solves most of the problems related to code deployment and versioning. Tools that were written for HPCs can be run on less powerful computers or vice versa and the OS is no longer a limiting factor for a program's dissemination. Using containers in the future will also improve the reliability of scientific experiments and allow for reproducibility. During peer review or after publication, entire datasets, programs, and results could be retrieved from repositories and tested thoroughly. In the field of protein annotation, containers allow wet-lab biologists to install server-grade pipelines through Biocontainers or GenomeHub (da Veiga Leprevost et al., 2017, Challis et al., 2017) that usually would take several days in a matter of minutes, complete with tools, code and biological databases. Stream-lined versions of these annotation tools can be installed for research groups or centers, without having to rely on an in-house or external dedicated bioinformatics service, drastically cutting the time and the costs required to analyze NGS data.

2 Datasets



2.1 PVC bacteria

The PVC superphylum is a grouping of distinct phyla in bacteria proposed on the basis of 16S rRNA gene sequence analysis (Wagner and Horn 2006). It initially consisted of a core of phyla *Planctomycetes*, *Verrucomicrobia* and *Chlamydiae*, but several other phyla have been considered to be members, including phylum *Lentisphaerae* (Cho et al., 2004), *Bacteroidetes* (Yutin et al., 2012), and potentially *Poribacteria* (Fieseler et al., 2004, Gupta et al., 2012). Several other phyla belonging to the PVC superphylum include only yet-to-be cultured members, like *Candidatus* Omnitrophica and *Kirimatiellaeota* (Rivas-Marin and Devos, 2018). While displaying diversity in shape, role, and functions, these phyla are all considered to be monophyletic (Wagner and Horn, 2006).

These organisms are remarkable for their complex cell biology, their evolutionary implications, and their role in the major biogeochemical cycles such as the carbon and nitrogen cycle (Strous et al., 1999, Lindsay et al., 2001, van Niftrik et al., 2004). Most of these traits make PVCs unique among bacteria (**Figure 8**).

PVC Superphylum

Features	Specific to	Found in
Compartmentalized cell plan (*20*)	Eu	Pl, Ve
DNA surrounded by membrane (*21*)	Eu	Pl
Condensed DNA (*22*)	Eu	Pl
Histone H1 (*23*)	Eu	Ch
Division by budding (*24*)***	Eu	Pl
Membrane coats (*11*)	Eu	Pl
Sterol (*25*)	Eu	Pl, Ch
Peptidoglycan loss (*26*)	Eu, Ar*	Pl, Ch
Proteic cell wall (*27*)	Eu	Pl
Ester and ether lipids (*28*)	Ar	Pl
FtsZ loss (*7*)	Eu, Ar**	Pl, Ch
Tubulin (*8, 9*)	Eu	Ve
C1 transfer (*29, 30*)	Ar	Pl
Endocytosis (*15*)	Eu	Pl

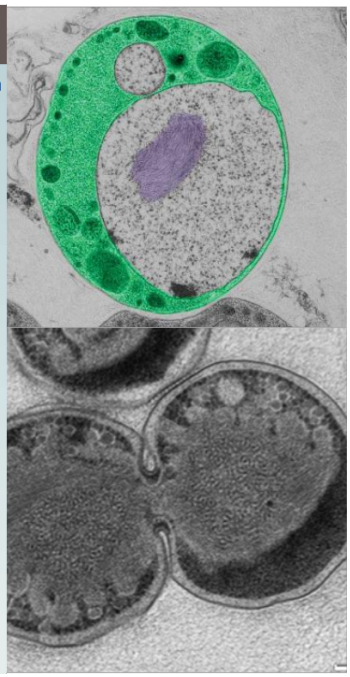


Figure 8. Unique traits of the PVC superphylum. The table on the left (from Devos and Reynaud, 2010) shows some of their features that aren't found in other bacteria. Eukaryotes (Eu), Archaea (Ar), *Planctomycetes* (Pl), *Chlamydiae* (Ch), *Verrucomicrobia* (Ve). The image on the top right (Rachel Melwig, EMBL) shows the complex cell plan of *G. obscuriglobus*. The bottom right image shows a member of the genus "*Candidatus* Kuenenia stuttgartiensis (L. van Niftrik, Radboud University).

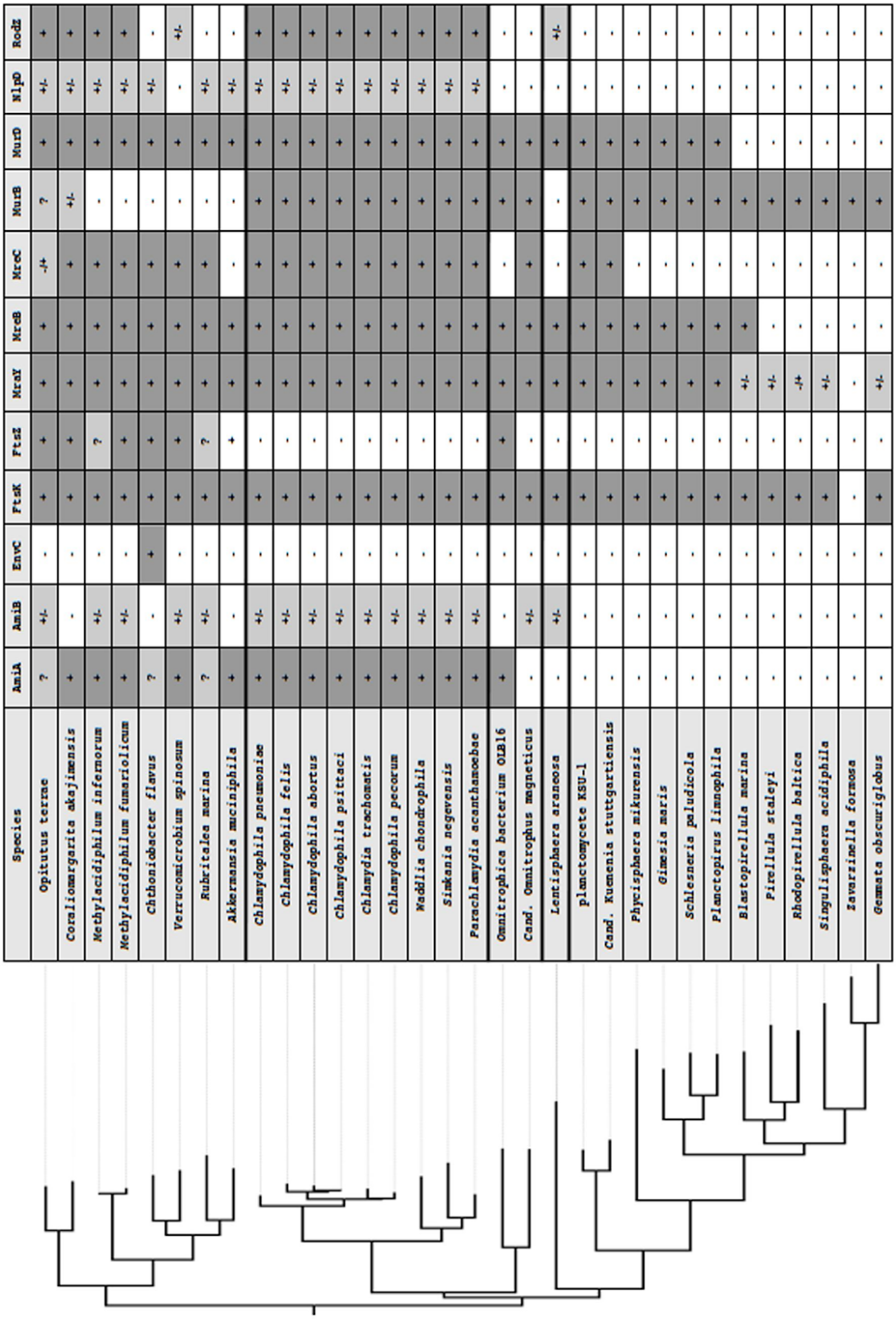
Among the members of the PVC superphylum, *Planctomycetes* is particularly interesting because they possess features that are uncommon in bacteria (Devos, 2013; Fuerst, 2013; Devos and Ward, 2014), some of which are more common in archaea or eukaryotes (Devos and Reynaud, 2010; Reynaud and Devos 2011). *Planctomycetes* shares a complex cell plan different from a typical Gram-negative (G-) bacteria, with their internal membranes have impressive intracytoplasmic invaginations (Santarella-Mellwig et al., 2013). Recent studies have shown that this cell plan is not different from a classical G-, but a variation (Devos, 2013a).

Their cytoplasmic membrane creates invaginations that cover most of the internal cell volume, sometimes almost engulfing the nucleoid, which is highly condensed (Yee et al., 2012). Other Planctomycetes have additional organelles like the anammoxosome, which is separated from the cytoplasm and devoted to nitrogen processing.

Although initially thought to lack peptidoglycan (PG), it has been recently reported that PG is present in at least 5 members of the *Planctomycetes* (Jeske et al., van Teeseling et al., 2015) and two *Chlamydiae* (Pilhofer et al., 2013; Liechti et al., 2014).

Some planctomycetal genomes contain genes coding for proteins that are structurally related to membrane coat proteins involved in the formation of the eukaryotic endomembrane system. Additionally, one of those bacterial proteins in *Gemmata obscuriglobus* was shown to be in close contact with the intracellular membrane (Santarella-Mellwig et al. 2010; Acehan et al. 2014). *G. obscuriglobus* can also intake macromolecules for their internal degradation, in a process that is similar to eukaryotic endocytosis (Lonhienne et al., 2010 Fuerst and Sagulenko 2014, Boedeker et al., 2017). Its membrane composition is quite unusual since it contains sterols, lipids found primarily in eukaryotes and only some bacteria (Pearson et al., 2003). Phylogenetic and biochemical evidence suggests that *G. obscuriglobus* contains the most ancient sterols synthesis pathway.

Another peculiar characteristic that separates *Planctomycetes* from other bacteria is their cell division mechanism. Most bacteria and a few Planctomycetes perform cell division through binary fission, while most members of the phylum divide through budding, similarly to *Saccharomyces cerevisiae*. They also lack the protein FtsZ, which is otherwise ubiquitous in bacteria and fundamental in cell division (**Figure 9**, Pilhofer et al., 2008 and Rivas-Marin et al., 2016).



Species	AmiA	AmiB	EnvC	FtsK	FtsZ	MraY	MreB	MreC	MreD	NlpD	RodZ
Opitutus terrae	?	+/-	-	+	+	+	+	-/+	?	+	+
Coraliomargarita akajimensis	+	-	-	+	+	+	+	+	+/-	+	+
Methylacidiphilum infernorum	+	+/-	-	+	?	+	+	+	-	+	+
Methylacidiphilum fumariolicum	+	+/-	-	+	+	+	+	+	-	+	+
Chthoniobacter flavus	?	-	+	+	+	+	+	+	-	+	-
Verrucomicrobium spinosum	+	+/-	-	+	+	+	+	+	-	-	+/-
Rubritalea marina	?	+/-	-	+	?	+	+	+	-	+/-	-
Akkermansia muciniphila	+	-	-	+	+	+	+	-	-	+/-	-
Chlamydophila pneumoniae	+	+/-	-	+	-	+	+	+	+	+/-	+
Chlamydophila felis	+	+/-	-	+	-	+	+	+	+	+/-	+
Chlamydophila abortus	+	+/-	-	+	-	+	+	+	+	+/-	+
Chlamydophila psittaci	+	+/-	-	+	-	+	+	+	+	+/-	+
Chlamydia trachomatis	+	+/-	-	+	-	+	+	+	+	+/-	+
Chlamydophila pecorum	+	+/-	-	+	-	+	+	+	+	+/-	+
Waddlia chondrophila	+	+/-	-	+	-	+	+	+	+	+/-	+
Simkania negevensis	+	+/-	-	+	-	+	+	+	+	+/-	+
Parachlamydia acanthamoebae	+	+/-	-	+	-	+	+	+	+	+/-	+
Omnitrophica bacterium OLB16	+	-	-	+	+	+	+	-	+	-	-
Cand. Omnitrophus magneticus	-	+/-	-	+	-	+	+	+	+	-	-
Lentisphaera araneosa	-	+/-	-	+	-	+	+	-	+	-	+/-
Planctomycete KSU-1	-	-	-	+	-	+	+	+	+	-	-
Cand. Kuenenia stuttgartiensis	-	-	-	+	-	+	+	+	+	-	-
Phycisphaera mikurensis	-	-	-	+	-	+	+	-	+	-	-
Gimesia maris	-	-	-	+	-	+	+	-	+	-	-
Schlesneria paludicola	-	-	-	+	-	+	+	-	+	-	-
Planctopirus limnophila	-	-	-	+	-	+	+	-	+	-	-
Blastopirellula marina	-	-	-	+	-	+/-	+	-	+	-	-
Pirellula staleyi	-	-	-	+	-	+/-	-	-	+	-	-
Rhodopirellula baltica	-	-	-	+	-	-/+	-	-	+	-	-
Singulisphaera acidiphila	-	-	-	+	-	+/-	-	-	+	-	-
Zavarzinella formosa	-	-	-	-	-	-	-	-	+	-	-
Gemmata obscuriglobus	-	-	-	+	-	+/-	-	-	+	-	-

Figure 9. Main genes involved in cell division and their distribution in PVC members (Source: Rivas-Marin et al., 2016).

These bacteria are also particularly relevant in the fields of ecology (Glöckner et al., 2003), evolution (González-Sánchez et al., 2015) and biotechnology (Devos and Ward, 2014). Five genera (Candidata Kuenenia, Brocadia, Anammoxoglobus, Jettenia and Scalindua), forming the order Brocadiales inside the phylum *Planctomycetes*, are currently used in wastewater treatment. They possess an anammoxosome that allows them to process ammonium anaerobically and several carbon-based sources (van Niftrik and Jetten, 2012). While *Planctomycetes* and *Verrucomicrobia* are present in almost every environment on Earth and in the human microbiome, it's worth remembering the relevance of *Chlamydiae* in modern medicine due to its infectious nature and ability to parasitize human beings.

The three main phyla, *Planctomycetes*, *Verrucomicrobia* and *Chlamydiae*, show an important variance in their proteome sizes. As a reference, some model bacteria like *Escherichia coli* or *Bacillus subtilis* have 4305 and 4197 different proteins respectively. *Chlamydiae* members have very reduced genomes with low protein numbers (mean/median: 1532/1125 proteins). *Chlamydia trachomatis* possesses one of the tiniest proteomes, with only 895 proteins. Verrucomicrobia appears to be intermediary (4307/4588), with some close to the size of reduced chlamydial pathogens with around 2000 proteins. In contrast, the Planctomycetes display much larger proteome sizes (5881/6193), which rank them amongst the bacteria with the biggest genomes and most protein-coding genes. Some of the largest PVC proteomes belong to the *Planctomycetaceae* family, with *Zavarzinella formosa*, *Rhodopirellula maiorica SM1* and *G. obscuriglobus* encoding 8123, 7825 and 7756 proteins respectively. These are almost one order of magnitude bigger than the smallest chlamydial proteome, *C. trachomatis*, and bigger than some eukaryotes like baker's yeast *S. cerevisiae* that encodes 6721 proteins.

Despite the considerable interest in these organisms, the number of uncharacterized proteins in their proteomes average at 46% (Bordin et al., 2018).

This problem is not uncommon with the current state of protein function prediction (See Introduction, Current state of Protein Function Prediction), since around 31% of all non-PVC sequences in UniProt are marked as "of unknown function", but it is particularly exacerbated in the case of PVC proteomes.

The case of the PVCs is interesting because knowledge about these organisms can vary greatly, from 18.6% of uncharacterized proteins in *Chlamydophila abortus* to the staggering amount of 88.6% in the case of *Rubritalea marina* (**Figure 10**).

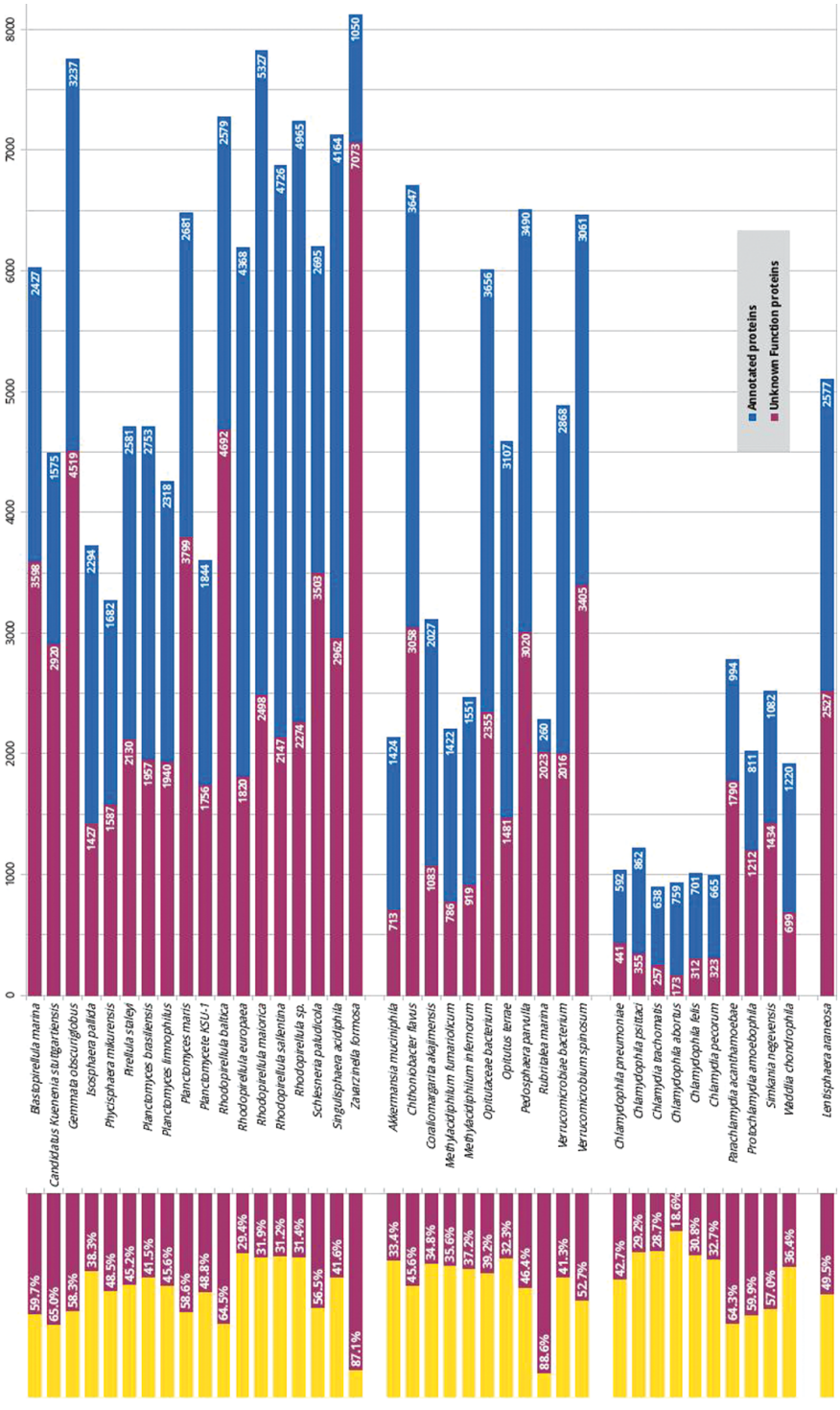


Figure 10 (Bordin et al., Database 2018) Current status of the PVC proteomes annotation.

Due to the interesting features of the PVC superphylum and its currently lacking level of annotation, we applied the concept of Integrative Cell Biology to them in order to improve their overall protein annotations.

2.2 The PVCbase dataset

The proteomes of 39 PVCs, comprising of 17 Planctomycetes, 11 Verrucomicrobia, 10 Chlamydiae and 1 Lentisphaerae, were retrieved from UniProt and the NCBI-protein databases (**Table 3**). In the case of *Planctomycetes* and *Verrucomicrobia*, the proteome of the representative strain for each species was selected, based on what was available or the most complete. For Chlamydias, several strains are completely sequenced but only one representative per species was selected in order to avoid redundancy.

Planctomycetes				
Organism	TaxID	Genome assembly	No. of proteins	Source
Blastopirellula marina DSM 3645	314230	Scaffold	6025	NCBI
Gemmata obscuriglobus UQM 2246	214688	Contig	7756	NCBI
Isosphaera pallida ATCC 43644	575540	Genome	3721	UniProt
Phycisphaera mikurensis NBRC 102666	1142394	Genome	3269	UniProt
Pirellula staleyi ATCC 27377	530564	Genome	4711	UniProt
Planctomyces brasiliensis ATCC 49424	756272	Genome	4710	UniProt
Planctomyces limnophilus DSM 3776	521674	Genome	4258	UniProt
Planctomyces maris DSM 8797	344747	Contig	6480	NCBI
Planctomycete KSU-1	247490	Contig	3600	NCBI
Rhodopirellula baltica SH 1	243090	Chrom.	7271	UniProt
Rhodopirellula europaea 6C	1263867	Contig	6188	NCBI
Rhodopirellula maiorica SM1	1265738	Contig	7825	NCBI
Rhodopirellula sallentina SM41	1263870	Contig	6873	NCBI
Rhodopirellula sp. SWK7	595460	Contig	7239	NCBI
Schlesneria paludicola DSM 18645	1123242	Scaffold	6198	NCBI
Singulisphaera acidiphila ATCC BAA-1392	886293	Genome	7126	UniProt
Zavarzinella formosa DSM 19928	1123508	Scaffold	8123	NCBI

Table 3. PVC Bacteria analyzed with the ICB pipeline available in the PVCdb section of PVCbase
(Bordin et al., Database 2018)

Verrucomicrobia				
Organism	TaxID	Genome assembly	No. of proteins	Source
Akkermansia muciniphila ATCC BAA-835	349741	Genome	2137	UniProt
Chthoniobacter flavus Ellin428	497964	Scaffold	6705	NCBI
Coraliomargarita akajimensis DSM 45221	583355	Genome	3110	UniProt
Methylacidiphilum fumariolicum SolV	1156937	Genome	2208	NCBI
Methylacidiphilum infernorum V4	481448	Genome	2470	UniProt
Opitutaceae bacterium TAV5	794903	Genome	6011	UniProt
Opitutus terrae PB90-1	452637	Genome	4588	UniProt
Pedosphaera parvulla Ellin514	320771	Contig	6510	NCBI
Rubritalea marina DSM 17716	1123070	Scaffold	2283	NCBI
Verrucomicrobiae bacterium DG1235	382464	Scaffold	4884	NCBI
Verrucomicrobium spinosum DSM 4136	240016	Genome	6466	NCBI

Chlamydiae				
Organism	TaxID	Genome assembly	No. of proteins	Source
Chlamydophila pneumoniae CWL029	115713	Genome	1033	NCBI
Chlamydophila psittaci 6BC	331636	Genome	1217	UniProt
Chlamydia trachomatis D/UW-3/CX	272561	Genome	895	UniProt
Chlamydophila abortus S26/3	218497	Genome	932	UniProt
Chlamydophila felis Fe/C-56	264202	Genome	1013	UniProt
Chlamydia pecorum E58	331635	Genome	988	UniProt
Parachlamydia acanthamoebae UV7	765952	Genome	2784	UniProt
Protochlamydia amoebophila UWE25	264201	Chrom.	2023	UniProt
Simkania negevensis Z	331113	Genome	2516	UniProt
Waddlia chondrophila ATCC VR-1470	716544	Genome	1919	UniProt

Table 3 (continued). PVC Bacteria analysed with the ICB pipeline available in the PVCdb section of PVCbase (Bordin et al., Database 2018)

Lentisphaerae				
Organism	TaxID	Genome assembly	No. of proteins	Source
Lentisphaera araneosa HTCC2155	313628	Contig	5104	NCBI

Table 3 (continued). PVC Bacteria analysed with the ICB pipeline available in the PVCdb section of PVCbase (Bordin et al., Database 2018)

This dataset comprises 173664 proteins, with a percentage of uncharacterized proteins of 49.07% for *Planctomycetes*, 44.28% for *Verrucomicrobia,* and 40.03% for *Chlamydiae*. The overall percentage for the complete dataset is 44.46%.

2.3 PVC bacteria attached to algae

Members of the PVC bacterial superphylum have been sampled in common and extreme habitats, ranging from topsoil to the depths of the ocean. Some of them live in symbiosis with other organisms, benefiting from the host interaction through the exchange of metabolites and the creation of biofilms.

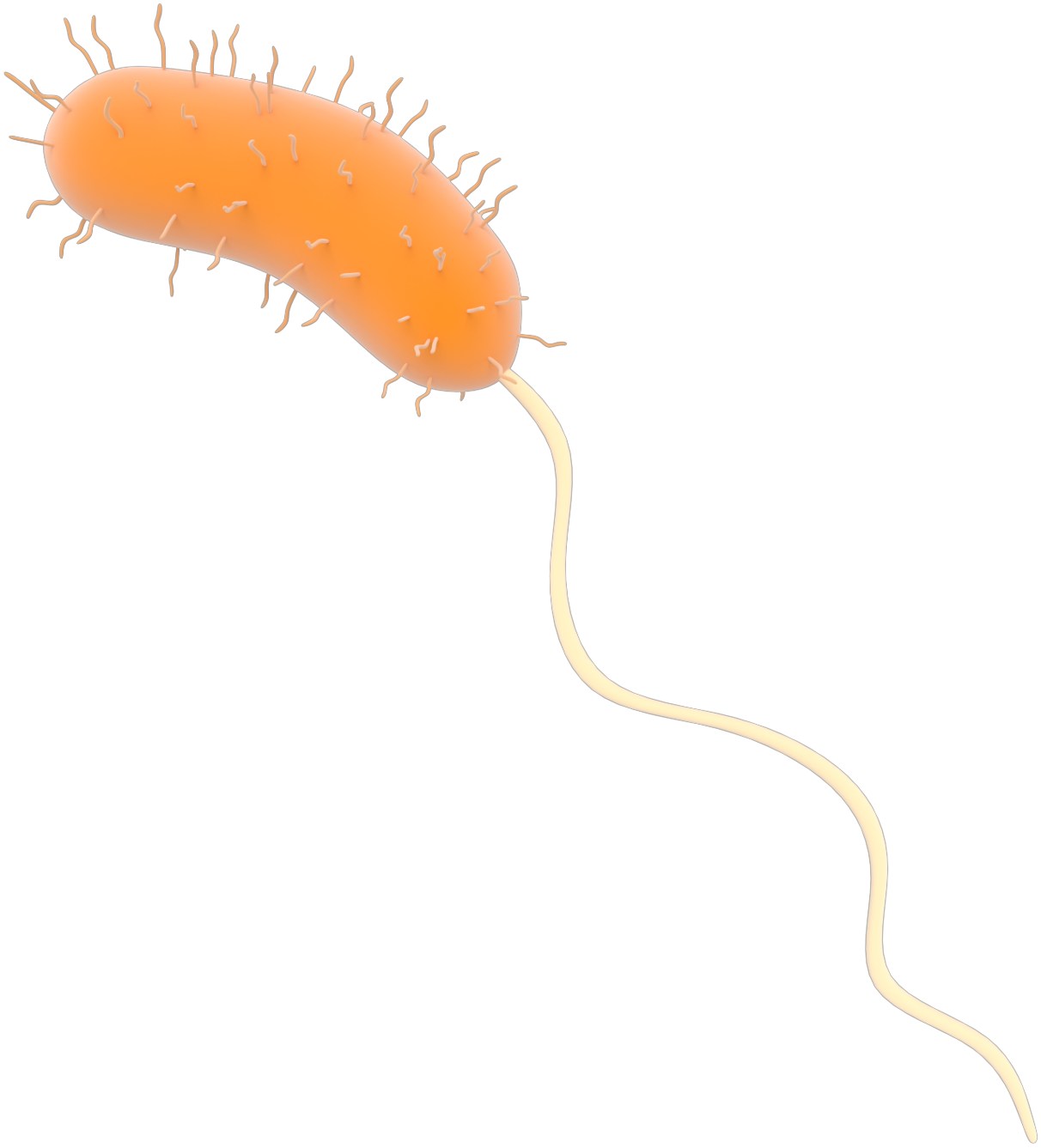
To further apply ICB to the PVCs, we additionally annotated three environmental *Planctomycetes* that were sampled from the surface of macroalgae (Lage and Bondoso 2011). These bacteria, *Roseimaritima ulvae* UC8, *Rubripirellula obstinata* LF1, and *Mariniblastus fucicola* FC18, live in a complex macroalgal biofilm found on *Ulva* sp., *Laminaria* sp., and *Fucus spiralis* along the northern coast of Portugal.

Besides trying to obtain a more complete annotation of these genomes, efforts were directed towards determining their shared genes, what differs them from Planctomycetes found in different environments, and the characterization of pathways that could explain their life in the macroalgae biofilm, the interaction with the algae, and the phylogenetic relationship with other PVCs.

Their genome size range is from 6.6 Mbp to 8.1 Mbp, encoding approximately 3500 to 4500 genes.

Attribute	Strains		
	LF1	UC8	FC18
Genome size (bp)	6,588,559	8,130,296	6,539,195
Contamination	1.16%	0.00%	0.11%
DNA GC content	54.1%	59.12%	53.4%
CDS (Prokka)	3958	4479	3543
tRNA genes	69	71	66
Contigs	309	108	64
ORFs	5200	5759	5096

Table 4. General overview of the genome features of LF1, UC8 and FC18.



3 Objectives

Objectives

The aim of this work is to transform the concept of Integrative Cell Biology for protein annotation into a viable set of tools and resources. The specific breakdown of the objectives is:

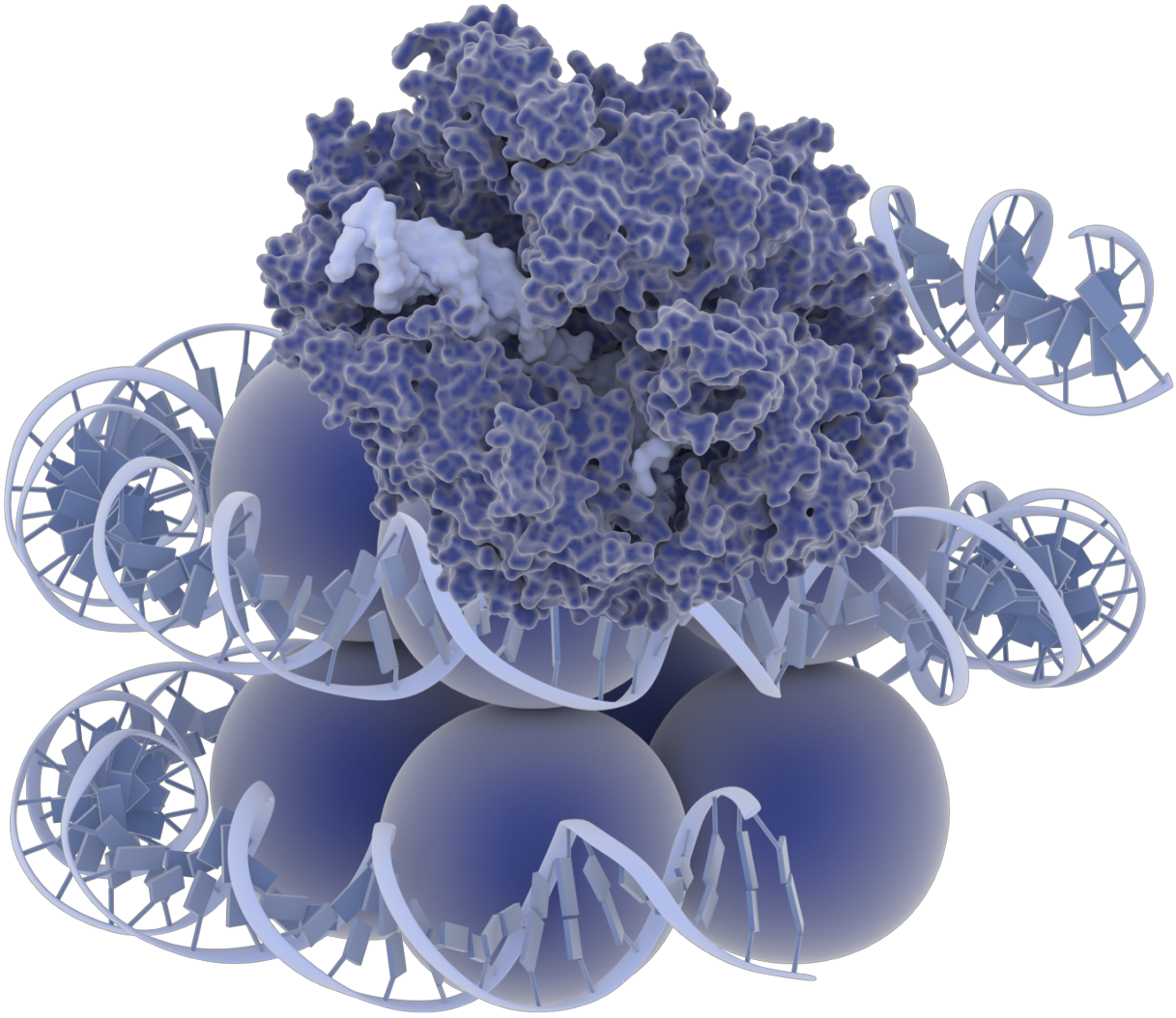
1. Creation of an ICB Computational Pipeline that integrates different protein predictors, organizes and presents the results in a user-friendly fashion.

2. Application of ICB to the proteomes of 42 PVC bacteria (39 PVCs and 3 Planctomycetes attached to algae), with the goal of improving the knowledge of these organisms. Measure the potential improvement in protein annotation within this group and use the raw data to further characterize the superphylum as a whole in comparison with other organisms.

3. Creation of PVCbase to show the results obtained through the annotation of the PVCs, including tools to allow further visual and sequence queries on their proteins. The goal is to make PVCbase the ultimate resource for PVC-associated research.

4. Create a Docker container for the ICB pipeline to allow users to easily download and obtain the predictions available for their organisms of interests.

4 Results



4.1

PVCbase: an integrated web resource for the PVC bacterial proteomes

DATABASE
The Journal of Biological Databases and Curation

Database, 2018, 1–10
doi: 10.1093/database/bay042
Original article

Original article

PVCbase: an integrated web resource for the PVC bacterial proteomes

Nicola Bordin[1,*], Juan Carlos González-Sánchez[2,3] and Damien P. Devos[1,*]

[1]Centro Andaluz de Biología del Desarrollo, CSIC, Universidad Pablo de Olavide, Carretera de Utrera, Km. 1, Seville 41013, Spain, [2]CellNetworks, BioQuant, University of Heidelberg, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany and [3]Biochemie Zentrum Heidelberg (BZH), Heidelberg University, Im Neuenheimer Feld 328, 69120 Heidelberg, Germany

*Corresponding author: Email: nbor1@upo.es

Correspondence may also be addressed to Damien P. Devos. Email: damienpdevos@gmail.com

Citation details: Bordin, N., González-Sánchez, J.C., and Devos, D.P. PVCbase: an integrated web resource for the PVC bacterial proteomes. *Database* (2018) Vol. 2018: article ID bay042; doi:10.1093/database/bay042

Received 26 January 2018; Revised 15 March 2018; Accepted 5 April 2018

Abstract

Interest in the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) bacterial superphylum is growing within the microbiology community. These organisms do not have a specialized web resource that gathers *in silico* predictions in an integrated fashion. Hence, we are providing the PVC community with PVCbase, a specialized web resource that gathers *in silico* predictions in an integrated fashion. PVCbase integrates protein function annotations obtained through sequence analysis and tertiary structure prediction for 39 representative PVC proteomes (PVCdb), a protein feature visualizer (Foundation) and a custom BLAST webserver (PVCBlast) that allows to retrieve the annotation of a hit directly from the DataTables. We display results from various predictors, encompassing most functional aspects, allowing users to have a more comprehensive overview of protein identities. Additionally, we illustrate how the application of PVCdb can be used to address biological questions from raw data.

Database URL: PVCbase is freely accessible at www.pvcbacteria.org/pvcbase

Introduction

The *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) bacterial superphylum is composed of the three name-giving phyla and some additional ones, like *Lentisphaerae*, '*Candidatus* Omnitrophica' and *Kirimatiellaeota* (1). Despite this diversity, it is now accepted that they form a monophyletic group (2). This bacterial superphylum draws interest because some species display characteristics not frequently observed in bacteria. Examples of these are condensed DNA (nucleoids), extensive inner membrane organization (3), the ability to internalize external compounds before degradation (4), the presence of membrane coat-like proteins linked to the extensive membrane organization (5, 6) and that they were thought to lack peptidoglycan

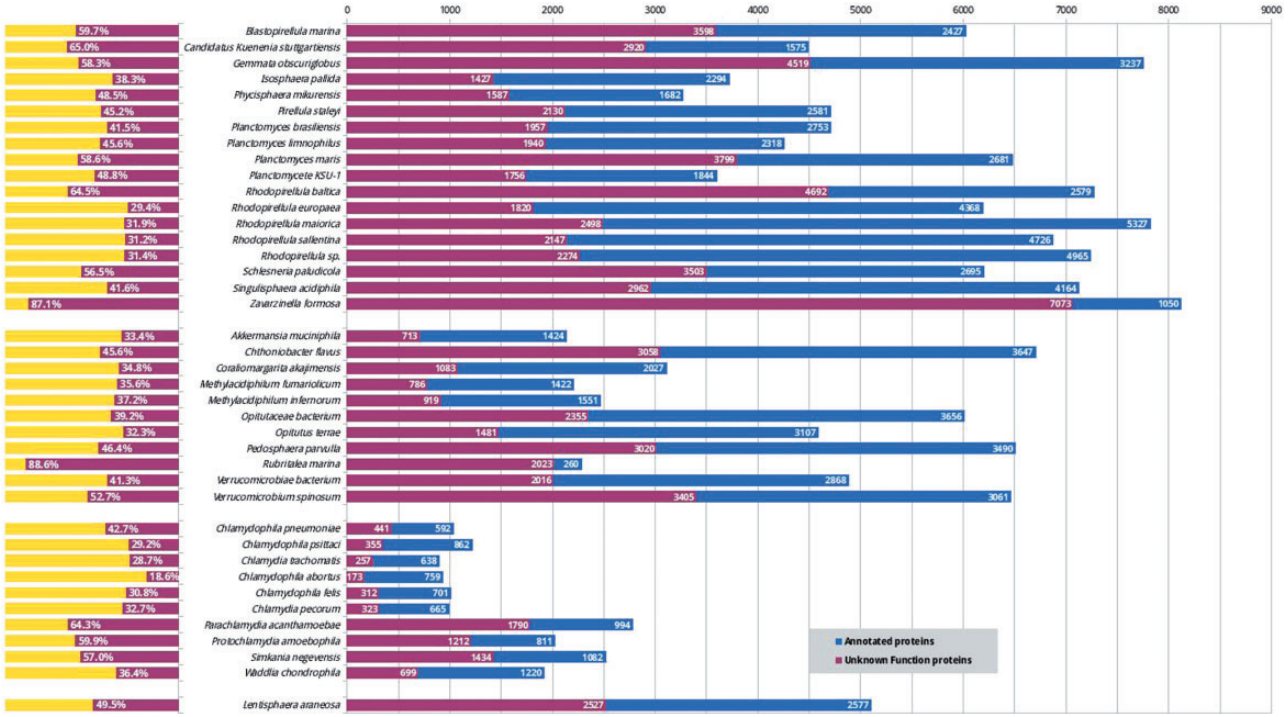


Figure 1. Status of functional annotation of PVC proteins. Total numbers (right) and percentage of total (left) proteins annotated (blue or yellow bars) and unannotated (purple) in each proteomes.

until recently (7, 8). Interesting characteristics relevant for wastewater treatment are shown by some planctomycetes having an anammoxosome, possibly the first true prokaryotic organelle, which allows the bacteria to degrade ammonium anaerobically (9). The exceptional diversity of cell plans displayed by some of the phyla, showcased by *Gemmata obscuriglobus* and some verrucomicrobia (10), has been the subject of controversial interpretations (11, 12).

Despite the considerable interest in these bacteria in many fields, including cell biology, evolution (13) and biotechnology (14), these organisms lack a centralized resource for their analysis. While new PVCs are being sequenced, the mean percentage of unannotated proteins constitutes approximately 46% (Figure 1). This issue does not exclusively affect the PVC bacteria since complete genome sequencing rarely translates into a complete characterization of the organism (in UniProt 31% of non-PVC proteomes are uncharacterized) (15). Protein function is not limited to a single feature or description; therefore we developed a pipeline for the simultaneous consideration of many different sequence descriptors. Since our aim is to provide these results to the community of PVC experimentalists, the results were collected in a resource built taking user-friendliness into consideration. Users are able to easily query the results of all functional predictors and download the predictions in bulk for large-scale interrogations.

Methods

Unless specified otherwise, all tools were used with default parameters.

Proteomes collection

The proteomes of 39 PVC representative species (17 planctomycetes, 11 verrucomicrobia, 10 chlamydia and 1 lentisphaerae) comprising a total of 173 664 protein sequences, were obtained from the UniProtKB and the NCBI-protein databases. The complete list of PVC species is given in supplementary (See online supplementary material for Table S1).

Homology-based inference

For every sequence, an homology search was performed using PSI-BLAST (16) with three iterations and default parameters, against the UniProtKB/Swiss-Prot database (release 2015_02) (17). The raw output was parsed to extract all the hits showing an *E*-value below 1E-3 and a minimum coverage of 75% of the query sequence. The first of these matches was selected as the best match and information regarding function under the form of GO description (18), keywords and enzymatic activity was assigned to the query protein. From the remaining matches, GO terms were extracted and counted, and they were reported only if they appeared in at least 10% of the hits.

Domain analysis

The tool InterProScan (v5.16-55.0) (19) was used to search for protein signatures by scanning their sequences against all its member databases: Pfam (release 28.0), TIGRFAM (release 15.0), PANTHER, ProSite (release 20.113), HAMAP (release 2015_11), PIRSF (release 3.01), Gene3D (release 3.5.0), SUPERFAMILY (release 1.75), PRINTS (release 42.0), SMART (release 6.2) and InterPro (20), using the parameters *-goterms* and *-pa*. Thanks to these options, entries from the InterPro database also provided functional information in the form of terms from GO and KEGG-pathway entries (21).

Tertiary structure prediction

A series of programs and utilities included in the HHsuite package (v2.0.16) were used. For each sequence, first HHblits (22) was used to construct a high-quality multiple sequence alignment (MSA) by comparing it against the UniProt20 database of template Hidden Markov models (HMMs) (release 2013_03), with the option *-addss* which adds secondary structure information predicted with PSIPRED v3.5 (23) to the resulting MSA. It was then converted to a HMM (.hhm format) with the *hhmake* function. Finally, HHsearch (24) was used to compare it against the HMM template database pdb70 (release 16May15), which is based on the protein data bank (PDB) (25). Every tool was run with default parameters. The results for both comparisons, against UniProt20 and pdb70, were parsed with an *E*-value cut-off of 1E-3. For the latter, functional information in the form of GO terms and EC codes was gathered from SIFTS mapping (26).

Prediction of signal peptides and transmembrane helices

Signal peptides were predicted with SignalP4.1 (27) using the *gram-* option. Transmembrane helices (TMHs) were predicted with TMHMM (28). The content (%) of integral membrane or transmembrane proteins of the proteome was defined as the fraction of proteins for which at least one TMH was predicted (transmembrane proteins/total number of proteins*100).

Prediction of protein intrinsic disorder

The IUPred tool (29) was used to predict intrinsically disordered regions and globular domains. The default threshold of 0.5 was used to determine whether a residue was considered as structured or disordered. Three metrics were computed to describe disorder within the proteome (30): (i) the disorder content (%) which was calculated as the fraction of disordered residues in the proteome (total predicted disordered residues/total number of residues*100); (ii) the content (%) of long disordered regions (LDRs), which are defined as those regions where at least 30 disordered residues are predicted continuously along the sequence, calculated as the fraction of residues in those LDRs, (residues in LDRs/total number of residues*100); and (iii) the fraction (%) of highly disordered proteins (HDPs) which are defined as those with >50% of predicted disordered residues in their sequences (number of HDP/total number of proteins*100).

PVCbase

PVCbase is a webserver developed to distribute the predictors results and statistics on disorder and TMHs distributions for the PVC superphylum. The resource acts as a gateway to PVCdb, the BLAST webserver and our secondary structure descriptor, Foundation. PVCbase is built on top of a Linux-Apache-MySQL-Python stack with WordPress as content management system (Figure 2).

PVCdb

PVCdb data include sequence and structure-based features that were computed using a Python-Perl pipeline that runs, parse and organize the predictors results (Figure 3), generating tabular and HTML web pages for each proteome. The HTML pages include Javascript code that retrieves the query originated from PVCBlast from the URI and prefilters the jQuery DataTable on load. The standard DataTables plugin was modified in order to allow fixed headers, table prefiltering and paging.

PVCBlast

PVCBlast is based on the SequenceServer (31) Ruby package running on a RubyOnRails-Passenger-Apache stack. The Ruby was customized to highlight hit significance, links to PVCbase, a sample FASTA file, further sequence checks and a link-out system that connects a hit on a PVC proteome to its annotation in PVCdb. We included some additional controls to the original SequenceServer, such as refresh and back-to-top buttons.

Foundation

For ease of visualization, we provide a protein features visualization tool, Foundation, that combines secondary structure (predicted by PSI-Pred), Transmembrane helices (by TMHMM) and disorder (by IUPred) predictions.

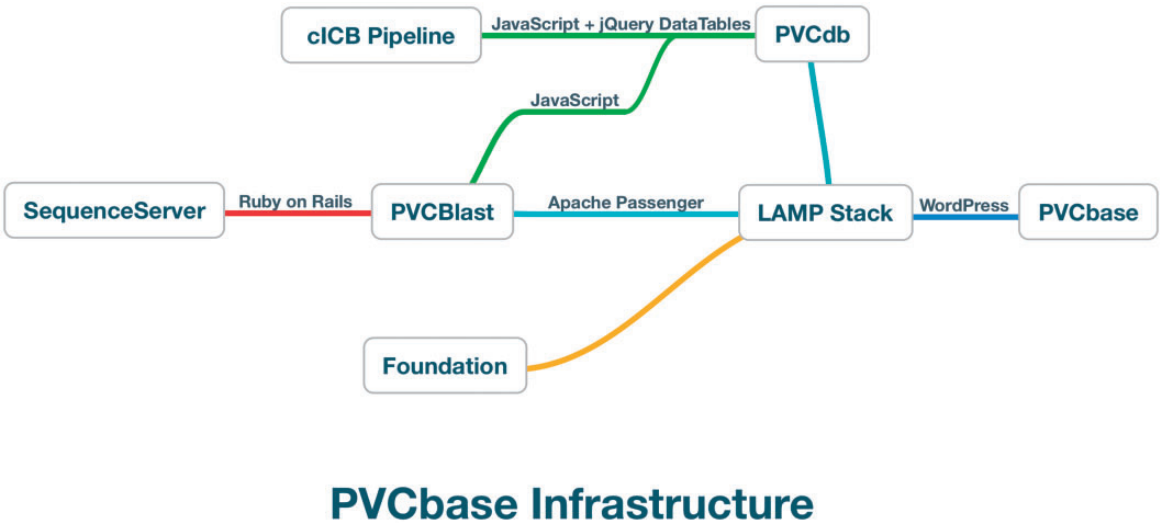


Figure 2. PVCbase organization and relationship between services.

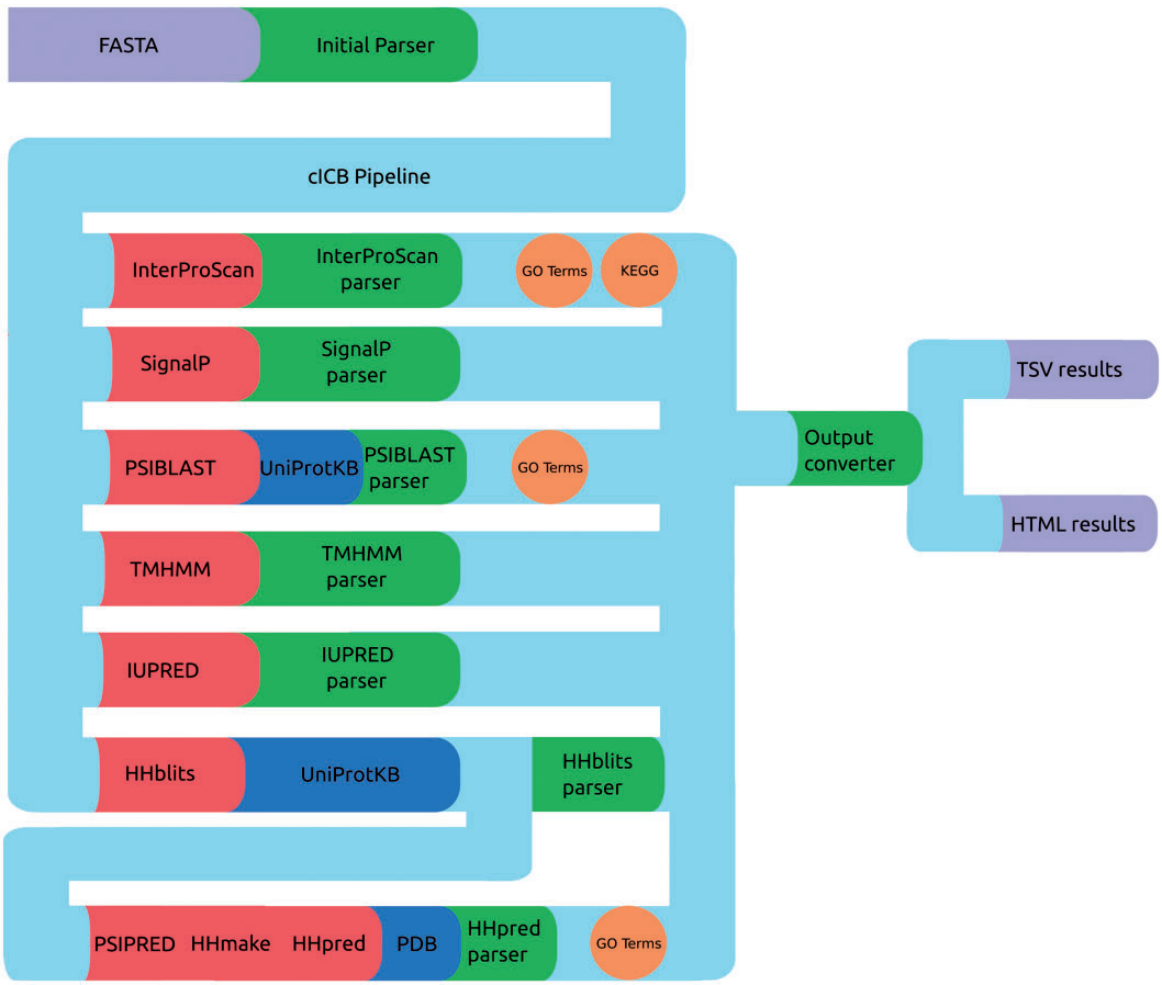


Figure 3. cICB pipeline. Files (purple), Tools (red), databases (blue) and scripts (green) used to generate the annotations for PVCdb. Color spheres indicates additional information obtained through the predictors.

In addition to a quick visualization of sequence feature, an illustration is provided in post-script, allowing the addition of more annotations, such as domains or mutated residues, as well as merging of various images, with limited understanding of postscript scripting. Foundation allows downloading of the raw results from the predictors for further analysis.

Results

PVC proteomes

We gathered the proteomes of 39 PVC species comprising 17 planctomycetes, 11 verrucomicrobia, 10 chlamydia and 1 lentisphaerae (See online supplementary material for Table S1). The three main phyla, *Planctomycetes, Verrucomicrobia* and *Chlamydia*, show an important variance of their proteome sizes. As reference, some of the model bacteria like *Escherichia coli* or *Bacillus subtilis* have 4305 and 4197 different proteins, respectively. Chlamydia has very reduced genomes with low protein numbers (mean/median: 1532–1125 proteins). *Chlamydia trachomatis* possesses one of the tiniest proteome, with only 895 proteins. The *Verrucomicrobia* appears to be intermediary (4307/4588) with some close to the size of reduced chlamydial pathogens with around 2000 proteins. In contrast, the *Planctomycetes* displays much larger proteome sizes (5881/6193), which rank them among the bacteria with the biggest genomes and most protein-coding genes. The largest PVC proteomes belong to the *Planctomycetaceae* family, with *Zavarzinella formosa*, *Rhodopirellula maiorica SM1* and *G.obscuriglobus* encoding 8123, 7825 and 7756 proteins, respectively, almost one order of magnitude bigger than the smallest chlamydial proteome, *C.trachomatis* and bigger than some eukaryotes, like the baker's yeast *Saccharomyces cerevisiae* that encodes 6721 proteins. The biggest planctomycetal genome, *Fimbriiglobus ruber*, has recently been reported to have a size of 12.364 Mbp and it encodes more than 10 000 proteins (32).

For comparison, one of the biggest bacterial genomes is found in *Ktedonobacter racemifer* in the phylum *Chloroflexi*, with 13.7 Mbp and coding for 11 000 proteins (33).

Usage of PVCbase

PVCdb

PVCdb collects protein functional annotations of 39 PVC proteomes. Each proteome can be downloaded as multiFASTA and the corresponding annotation can be either downloaded as a tabular file or easily browsed online. PVCdbs can show a variable amount of entries, based on user choice. Table rows can be sorted by length or alphabetically, while the search bar filters the table and shows only hits that contain the searched keyword. This is helpful to extract subsets of proteins based on localization, process or related to specific activities (Figure 4).

PVCBlast

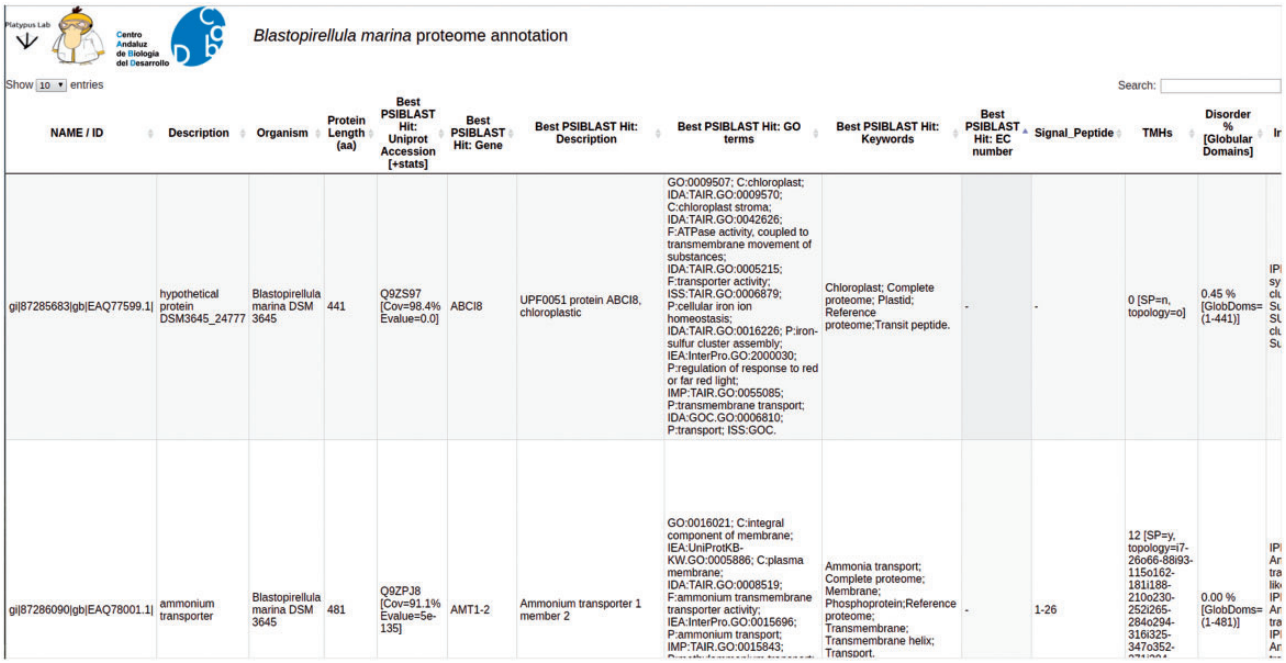
PVCBlast allows the user to perform BLAST searches on the PVC proteomes and genomes. The search box supports drag-and-drop and multiple sequences at once. The BLAST search parameters, such as evalues cutoff and number of alignments, can be customized using the 'advanced parameters' bar at the bottom of the page. The results page shows the alignments produced by BLAST, alongside several options for downloading the hits sequences and reports from the tool. The default SequenceServer interface provides a link-out service to NCBI Genbank or UniProt, according to the proteome source. We modified the aligned hits window to indicate the significance of a hit using shades of red. A link-out generator was created to link a hit on a PVC proteome to the corresponding annotation in PVCdb, pre-filtering the DataTable (Figure 5).

Foundation

Foundation is a tool to quickly visualize the linear and secondary structure features for a provided protein with its secondary structure features. Results can be downloaded as a png or postscript file, as a compressed tar file containing the raw output of the predictors, or can be visualized as an interactive zooming map. Secondary structure features are depicted with different color bars, fuchsia for α helices and cyan for β sheets. TMHs are shown as green boxes and the line underneath the secondary structure shows the disorder level for each amino acid (Figure 6).

Transmembrane proteins, intrinsic disorder and internal membrane: applications of PVCdb

In order to illustrate one of the possible discoveries made possible by PVCbase that would have been difficult to realize with other currently existing databases, we provide the following example. Planctomycetes cells present internal organizations characterized by the presence of developed membrane organizations which are atypical for bacteria (12). These have been extensively studied in *G.obscuriglobus* by three-dimensional tomography reconstructions that reported an extensive network of internal membranes (3, 6). A separate compartment has also been described in anammox planctomycetes (34). Variations of



NAME / ID	Description	Organism	Protein Length (aa)	Best PSIBLAST Hit: Uniprot Accession [+stats]	Best PSIBLAST Hit: Gene	Best PSIBLAST Hit: Description	Best PSIBLAST Hit: GO terms	Best PSIBLAST Hit: Keywords	Best PSIBLAST Hit: EC number	Signal_Peptide	TMHs	Disorder % [Globular Domains]	Ir
gi|87285683|gb|EAQ77599.1|	hypothetical protein DSM3645_24777	Blastopirellula marina DSM 3645	441	Q9ZS97 [Cov=98.4% Evalue=0.0]	ABCI8	UPF0051 protein ABCI8, chloroplastic	GO:0009507; C:chloroplast; IDA:TAIR.GO:0009570; C:chloroplast stroma; IDA:TAIR.GO:0042626; F:ATPase activity, coupled to transmembrane movement of substances; IDA:TAIR.GO:0005215; F:transporter activity; ISS:TAIR.GO:0006879; P:cellular iron ion homeostasis; IDA:TAIR.GO:0016226; P:iron-sulfur cluster assembly; IEA:InterPro.GO:2000030; P:regulation of response to red or far red light; IMP:TAIR.GO:0055085; P:transmembrane transport; IDA:GOC.GO:0006810; P:transport; ISS:GOC.	Chloroplast; Complete proteome; Plastid; Reference proteome;Transit peptide.	-	-	0 [SP=n, topology=o]	0.45 % [GlobDoms= (1-441)]	IP sy cl Su Su cl Su
gi|87286090|gb|EAQ78001.1|	ammonium transporter	Blastopirellula marina DSM 3645	481	Q9ZPJ8 [Cov=91.1% Evalue=5e-135]	AMT1-2	Ammonium transporter 1 member 2	GO:0016021; C:integral component of membrane; IEA:UniProtKB-KW.GO:0005886; C:plasma membrane; IDA:TAIR.GO:0008519; F:ammonium transmembrane transporter activity; IEA:InterPro.GO:0015696; P:ammonium transport; IMP:TAIR.GO:0015843;	Ammonia transport; Complete proteome; Membrane; Phosphoprotein;Reference proteome; Transmembrane; Transmembrane helix; Transport.	-	1-26	12 [SP=y, topology=i7-26o66-88i93-115o162-181i188-210o230-252i265-284o294-316i325-347o352-	0.00 % [GlobDoms= (1-481)]	IP Ar tra lik IP Ar tra IP Ar

Figure 4. PVCdb. Example of Web interface of PVCdb. DataTable of Blastopirellula marina. The figure shows portion of the bacterial annotation. Each column can be sorted and browsed using the built-in search bar.

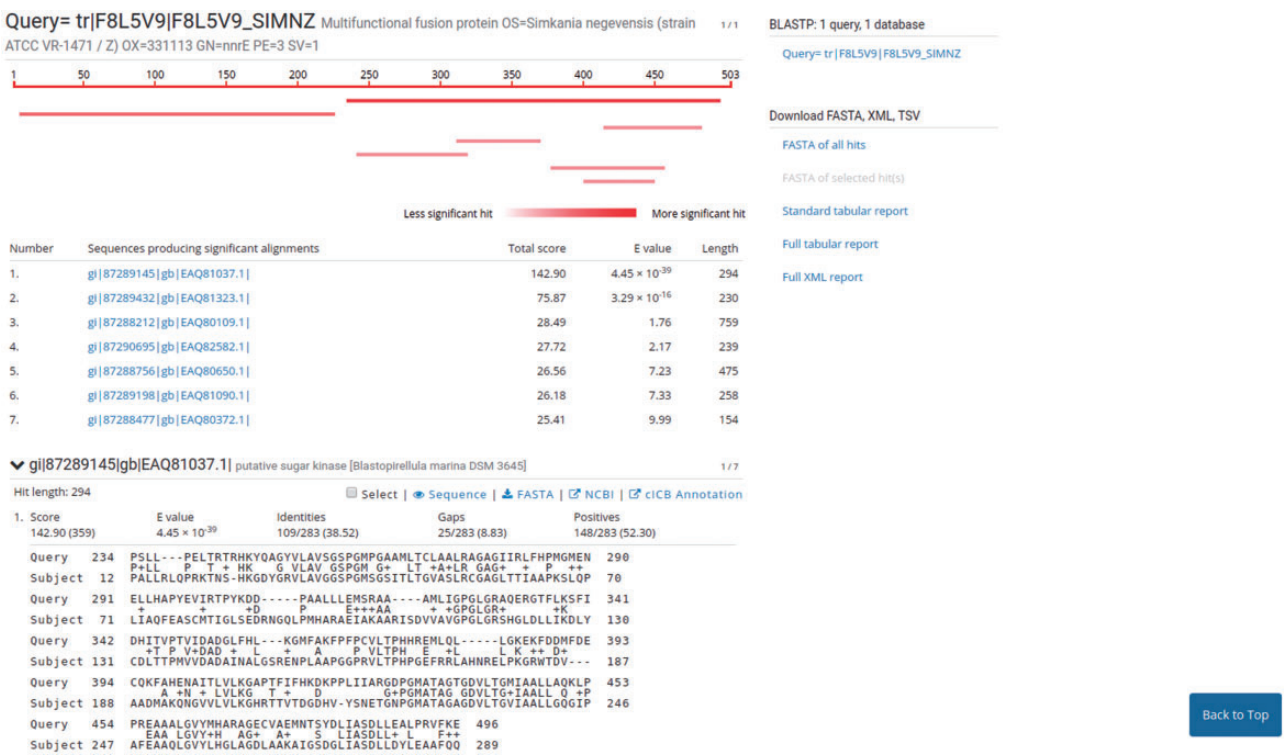


Figure 5. Results page of PVCBlast. The results page of PVCBlast provides the user with the alignments and the ability to download the results in various formats (top right). Among the linkout systems (center), clicking on 'cICB annotation' opens the correspondent annotation on PVCdb.

this cellular structure are believed to be shared by other planctomycetes and verrucomicrobia but are still mostly unexplored (10, 12, 35). We investigated a possible relationship between the presence of internal membrane and the number of transmembrane proteins. We calculated the fraction of proteins with TMHs for each PVC proteome and for each proteome of three reference sets composed of non-PVC bacterial, archaeal and eukaryotic

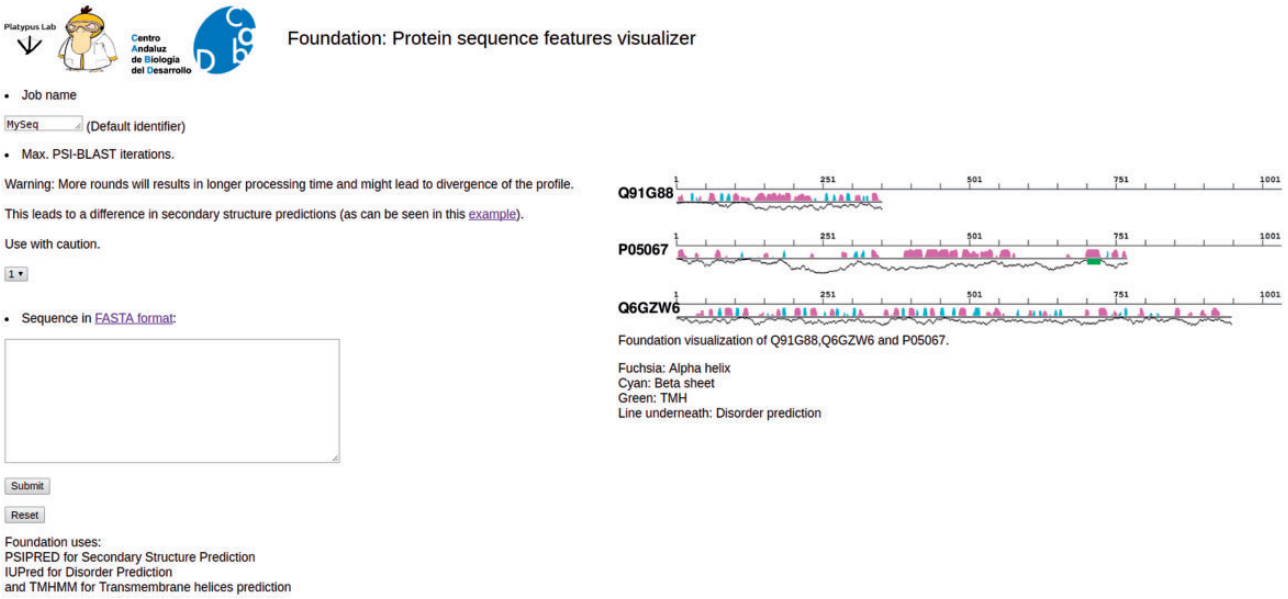


Figure 6. Foundation. Snapshot of Foundation's main page. The input box allows only single protein sequences. The number of PSIBLAST iterations can be modified using the drop-down menu (center-left). On right side of the page there are some examples of the output and the legend.

species (see Methods). These reference proteomes are representative of the three domains of life with diverse cellular plans (See online supplementary material for Table S2).

We first noticed that bacteria, in general, show a slightly higher content of TMHs than the analyzed species of archaea (*P*-value = 5.00E-04) and Eukaryotes (*P*-value = 6.92E-04). The content in transmembrane proteins showed however no statistical difference for any of the PVC bacterial groups when compared against all other groups and assessed by the Wilcoxon rank-sum test (See online supplementary material for Table S5, Figure 7). Therefore, the PVC genomes possess a smaller fraction of transmembrane protein in comparison to other bacteria (all means and medians between 23 and 24%). Additionally, we compared the fraction of transmembrane proteins to their number of TMHs (See online supplementary material for Table S3). The results provided further evidence for the previous observation. Thus, transmembrane protein content does not seem to be correlated with membrane complexity. This is illustrated by comparing the human and *C.trachomatis* proteomes, which respectively contain 17.41% and 24.58% of proteins containing at least one TMH.

Similarly, we explored the differences in protein structural disorder between these groups. We computed three metrics to describe the intrinsic disorder of the proteomes: the total disorder content, the fraction of residues in LDRs and the fraction of highly disordered proteins (HDPs) (see Methods) (Figure 8; See online supplementary material for Table S4 . As demonstrated elsewhere, disorder content is, for both prokaryotes and eukaryotes, generally independent of the proteome size (36). This observation is also reflected in our data. It is worth noting that the two outliers in the data belonging to archaea, and the largest one from the non-PVC bacteria values, correspond to three extreme halophilic organisms (the archaea *Halobacterium salinarum NRC1* and *Nanosalina sp*., and the bacterium *Salinibacter ruber*) (See online supplementary material for Figures S1 and S2). This observation agrees with the suggestion that intrinsically disordered proteins may help these organisms adapt to the extreme environments they inhabit. Disordered regions have an increased tolerance against mutations which allows for a higher evolutionary rate that results in extraordinary adaptability (37).

We first observed, as previously reported, that eukaryotes have more disordered proteomes (10), which has been related to the importance of disorder for cellular complexity. We then detected that most planctomycetes show a higher content of disordered proteins (mean/median of 12.15–12.39%) than non-PVC bacteria (6.82–5.95%). This trend is not observed in verrucomicrobia (7.43–7.08%) or chlamydia (4.95–5.14%). A statistical evaluation confirmed both observations: the disorder contents of planctomycetes species are significantly superior to the other three bacterial groups (Wilcoxon rank-sum test *P*-values: P vs V = 7.49E-04, P vs C = 6.64E-04, P vs non-PVC = 2.97E-04) and that the values from the other groups are not statistically different from each other (See online supplementary material for Table S5).

Low complexity, or intrinsically disordered proteins, ise mostly associated with signal transduction, cell-cycle regulation and transcription (38). However, it has been

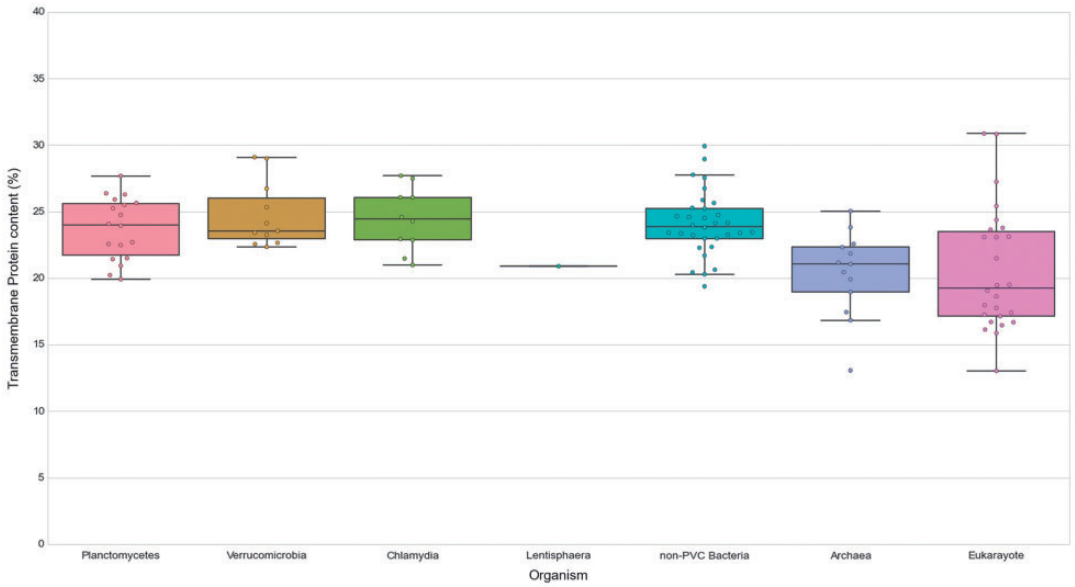


Figure 7. Transmembrane proteins in PVC and representative species from the tree of life. The numbers of TMHs containing proteins, expressed as percentages of the proteomes. Box plots reflect the distribution of the data. The box encloses the quartiles of the dataset, while the whiskers extend to the limits of the distributions. Outliers are determined based on the interquartile range and are not included in the boxes. The middle horizontal line in the box marks the median of the distribution.

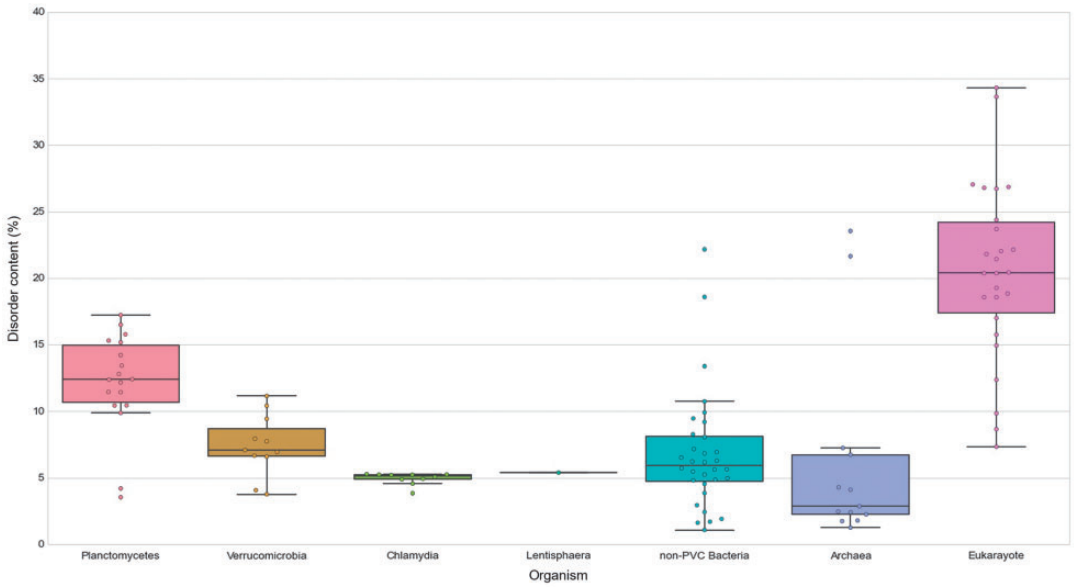


Figure 8. Disorder content in PVC and representative species from the tree of life. The numbers of disordered proteins, expressed as percentages of the proteomes. Box plots reflect the distribution of the data. The box encloses the quartiles of the dataset while the whiskers extend to the limits of the distributions. Outliers are determined based on the interquartile range and are not included in the boxes. The middle horizontal line in the box marks the median of the distribution.

suggested that structural disorder plays a fundamental role in vesicle trafficking pathways (39) and it has been demonstrated that certain unstructured protein domains are highly efficient drivers of membrane curvature. Disordered fragments have a role in membrane coat assembly and vesicle communication, in what has been called the fly-casting mechanism (40). This is especially the case in the clathrin-coated vesicle system, which mediates endocytosis and the early secretory pathway (35). Thus our observation suggests that the significantly higher ratio of disordered proteins in *Planctomycetes* appears to be correlated with the development of their membranes.

Conclusions

PVCbase offers a convenient one-stop platform for the PVC bacteria community. Its scalability and variety of annotations have already been used in PVC-related publications (41) and newly sequenced organisms will be added regularly. Bulk data collecting also allows users to infer biological discoveries by comparing annotations at the proteome level.

Availability

PVCbase is freely accessible at http://pvcbacteria.org/pvcbase.

Acknowledgements

We acknowledge Caitlin Lee Carpenter for her help in proofreading this article.

Supplementary data

Supplementary data are available at *Database* Online.

Funding

N.B. was funded by Marie Curie ITN FP7-ITN316723-PerFuMe. D.P.D. and J.C.G.S. were funded by the C2A grant EE: 2013/2506 from the Andalusian government. DPD was funded by the Spanish Ministry of Economy and Competitiveness (Grant nos. BFU2013-40866-P and BFU2016-78326-P).

Conflict of interest. None declared.

References

1. Rivas-Marín,E. and Devos,D.P. (2017) The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie Leeuwenhoek* doi: 10.1007/s10482-017-0962-z. [Epub ahead of print]
2. Wagner,M. and Horn,M. (2006) The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.*, **17**, 241–249.
3. Santarella-Mellwig,R., Pruggnaller,S., Roos,N. *et al.* (2013) Three-dimensional reconstruction of bacteria with a complex endomembrane system. *PLoS Biol.*, **11**, e1001565.
4. Lonhienne,T.G.A., Sagulenko,E., Webb,R.I. *et al.* (2010) Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus*. *Proc. Natl. Acad. Sci. U.S.A*, **107**, 12883–12888.
5. Santarella-Mellwig,R., Franke,J., Jaedicke,A. *et al.* (2010) The compartmentalized bacteria of the Planctomycetes-Verrucomicrobia-Chlamydiae superphylum have membrane coat-like proteins. *PLoS Biol.*, **8**, e1000281.
6. Acehan,D., Santarella-Mellwig,R. and Devos,D.P. (2014) A bacterial tubulovesicular network. *J. Cell. Sci.*, **127**, 277–280.
7. Jeske,O., Schüler,M., Schumann,P. *et al.* (2015) Planctomycetes do possess a peptidoglycan cell wall. *Nat. Commun.*, **6**, 7116.
8. van Teeseling,M.C.F., Mesman,R.J., Kuru,E. *et al.* (2015) Anammox planctomycetes have a peptidoglycan cell wall. *Nat. Commun.*, **6**, 6878.
9. van Niftrik,L. (2013) Cell biology of unique anammox bacteria that contain an energy conserving prokaryotic organelle. *Antonie Leeuwenhoek*, **104**, 489–497.
10. Lee,K.-C., Webb,R.I., Janssen,P.H. *et al.* (2009) Phylum verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum planctomycetes. *BMC Microbiol.*, **9**, 5.
11. Fuerst,J.A. (2013) The PVC superphylum: exceptions to the bacterial definition? *Antonie Leeuwenhoek*, **104**, 451–466.
12. Devos,D.P. (2014) PVC bacteria: variation of, but not exception to, the gram-negative cell plan. *Trends Microbiol.*, **22**, 14–20.
13. González-Sánchez,J.C., Costa,R. and Devos,D.P. (2015) A multi-functional tubulovesicular network as the ancestral eukaryotic endomembrane system. *Biology (Basel)*, **4**, 264–281.
14. Devos,D.P. and Ward,N.L. (2014) Mind the PVCs. *Environ. Microbiol.*, **16**, 1217–1221.
15. Galperin,M.Y. and Koonin,E.V. (2010) From complete genome sequence to "complete" understanding? *Trends Biotechnol.*, **28**, 398–406.
16. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Bairoch,A., Boeckmann,B., Ferro,S. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.
18. Ashburner,M., Ball,C.A., Blake,J.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
19. Jones,P., Binns,D., Chang,H.-Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
20. Hunter,S., Apweiler,R., Attwood,T.K. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
21. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
22. Remmert,M., Biegert,A., Hauser,A. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
23. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
24. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
25. Berman,H.M., Battistuz,T., Bhat,T.N. *et al.* (2002) The protein data bank. *Acta Crystallogr D Biol. Crystallogr.*, **58**, 899–907.
26. Velankar,S., Dana,J.M., Jacobsen,J. *et al.* (2012) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
27. Petersen,T.N., Brunak,S., von Heijne,G. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.
28. Krogh,A., Larsson,B., von Heijne,G. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

29. Dosztányi,Z., Csizmok,V., Tompa,P. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
30. Pietrosemoli,N., Pancsa,R. and Tompa,P. (2013) Structural disorder provides increased adaptability for vesicle trafficking pathways. *PLoS Comput. Biol.*, **9**, e1003144.
31. Priyam,A., Woodcroft,B.J., Rai,V. *et al.* (2015). SequenceServer: a modern graphical user interface for custom BLAST databases. biorxiv doi: 10.1101/033142.
32. Kulichevskaya,I.S., Ivanova,A.A., Baulina,O.I. *et al.* (2017) Fimbriiglobus ruber gen. nov., sp. nov., a Gemmata-like planctomycete from Sphagnum peat bog and the proposal of Gemmataceae fam. *Nov. Int. J. Syst. Evol. Microbiol.*, **67**, 218–224. 2017 Feb;
33. Chang,Y-j., Land,M., Hauser,L. *et al.* (2011) Non-contiguous finished genome sequence and contextual data of the filamentous soil bacterium Ktedonobacter racemifer type strain (SOSP1-21). *Stand. Genomic Sci.*, **5**, 97–111.
34. Neumann,S., Wessels,H.J.C.T., Rijpstra,W.I.C. *et al.* (2014) Isolation and characterization of a prokaryotic cell organelle from the anammox bacterium Kuenenia stuttgartiensis. *Mol. Microbiol.*, **94**, 794–802.
35. Boedeker,C., Schüler,M., Reintjes,G. *et al.* (2017) Determining the bacterial cell biology of Planctomycetes. *Nat. Commun.*, **8**, 14853.
36. Xue,B., Dunker,A.K. and Uversky,V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.*, **30**, 137–149.
37. Xue,B., Williams,R.W., Oldfield,C.J. *et al.* (2010) Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst. Biol.*, **4**, S1.
38. Tantos,A., Han,K.-H. and Tompa,P. (2012) Intrinsic disorder in cell signaling and gene transcription. *Mol. Cell. Endocrinol.*, **348**, 457–465.
39. Busch,D.J., Houser,J.R., Hayden,C.C. *et al.* (2015) Intrinsically disordered proteins drive membrane curvature. *Nat. Commun.*, **6**, 7875.
40. Shoemaker,B.A., Portman,J.J. and Wolynes,P.G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A*, **97**, 8868–8873.
41. Faria,M., Bordin,N. *et al.* (2017) Planctomycetes attached to algal surfaces: insight into their genomes. *Genomics*, pii: S0888-7543(17)30135-0.

Supplementary information and results from PVCbase: an integrated web resource for the PVC bacterial proteomes

We used three methods that provide a direct functional assignment: PSI-BLAST for the detection of homolog proteins, InterProScan for detecting protein signatures (regions and domains associated with specific protein families and functions) and HHPred for the detection of related proteins with known structures. However, all of them provide different types of information, such as the function of the most direct homologue, its E.C. number, and the main GO associated to the top 10 hits.

All reported results met our coverage and e-value thresholds. The results from each of the three methods were: The PSI-BLAST module detected homologues for 75,733 proteins (43.6% of the total); InterProScan found at least one functionally relevant protein signature for 121,128 proteins (69.8%); Tertiary structure models from the PDB were found with HHpred for 119,525 proteins (68.8%). Finally, 73,506 proteins (42.3% of the total) were annotated by the three main methods at the same time. 51,785 proteins of the total of 173644 (29.9%) didn't have either a PSI-BLAST or InterPro prediction. Of those, 7,566 (14.6%) do have a PDB hit. For those proteins, a full atom 3D model and functional information could then be derived from structurally related proteins.

Breakdown of the improved annotations obtained through ICB

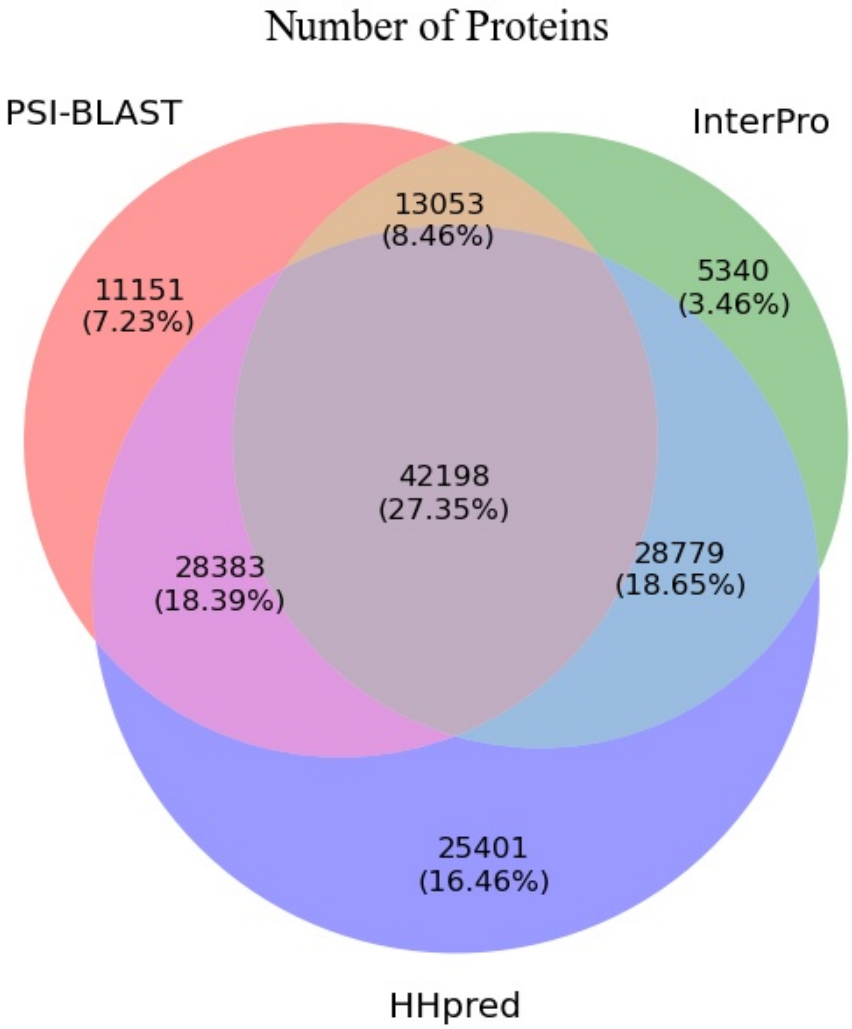


Figure 11. Number of new annotations divided by method.

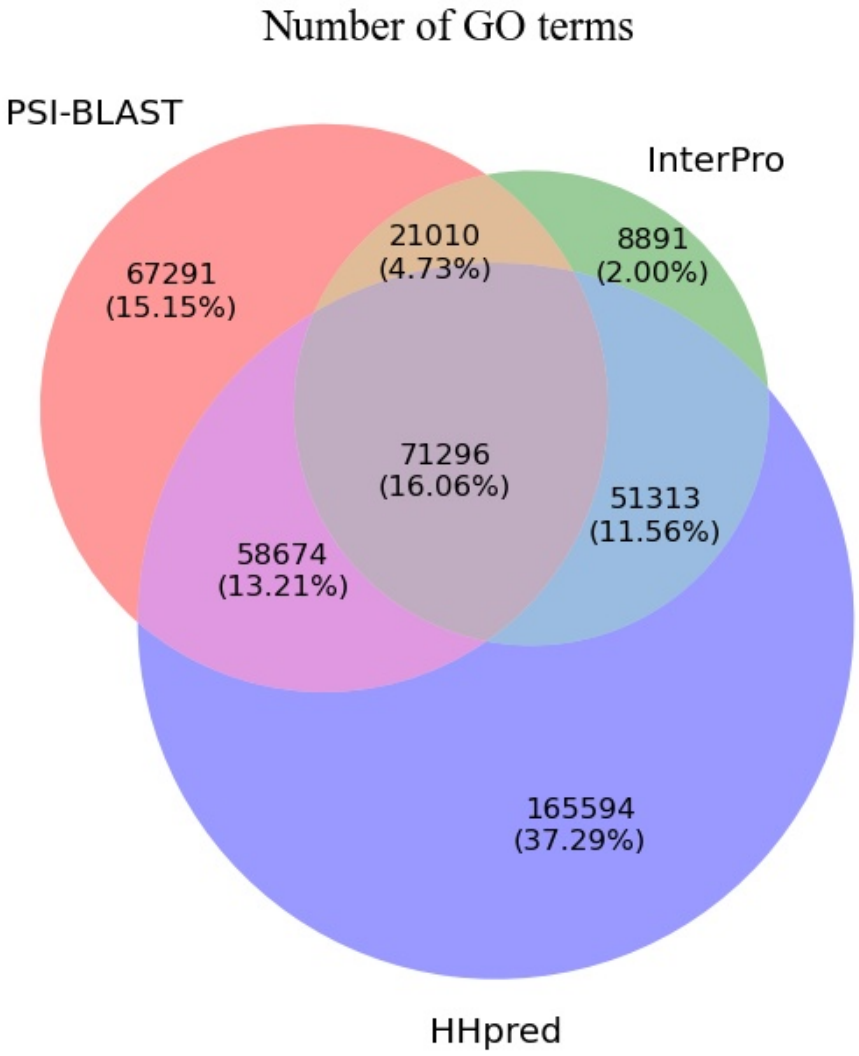


Figure 12. Number of GOs predicted through each method.

Primary structure features

Additionally, we described some primary features predictions, such as signal peptide at the N-terminus of the sequence, the number of transmembrane helices (TMHs) and their topology, and globularity reported as the fraction of predicted disordered residues.

Determining primary sequence features, such as localization or interactions, could provide key information for proteins that cannot be annotated by homology. For example, 11070 proteins in the dataset (25%) have at least 1 predicted TMH. The function of these proteins is unknown, but we can predict that they are located in the membrane.

Most proteins with predicted TMH, and thus likely targeted to the membrane, also have signal peptides. This data provided additional information that complemented the functional annotations obtained by the other modules. For example, the prediction of a secreted protein could be corroborated by the prediction of a signal peptide. The prediction of a long globular region in the protein (and therefore, a low disorder content) would agree with the assignment of a fold in that sequence region.

Table S2. Non-PVC organisms used for disorder and TMHs content comparisons.

Organism	Tax ID	Genome Assembly Level	No. of proteins
Bacteria (excluding PVC)			
Agrobacterium fabrum C58	176299	Complete Genome	5344
Azotobacter vinelandii	354	Contig	4990
Bacillus subtilis	1423	Contig	4197
Caulobacter crescentus	190650	Complete Genome	3720
E coli K12	83333	Complete Genome	4306
Lactobacillus casei	1582	Complete Genome	2708
Magnetospirillum magneticum	84159	Complete Genome	4514
Mycoplasma genitalium	243273	Complete Genome	483
Pseudomonas aeruginosa	287	Contig	5563
Pseudomonas fluorescens	216595	Complete Genome	6388
Salmonella typhimurium	90371	Contig	4533
Shigella dysenteriae	984897	Complete Genome	3897
Staphylococcus aureus	93061	Contig	2889
Yersinia pestis	632	Contig	3909
Aquifex aeolicus VF5	63373	Complete Genome	1553
Bifidobacterium longum	216816	Scaffold	1725
Chloroflexus aurantiacus	1108	Complete Genome	3850
Corynebacterium glutamicum	1718	Complete Genome	3093
Dictyoglomus turgidum	513050	Complete Genome	1743
Fusobacterium nucleatum	851	Complete Genome	2046
Gemmatimonas aurantiaca	173480	Complete Genome	3932
Leptospira interrogans	173	Contig	3676
Mesoplasma florum	2151	Complete Genome	683
Moorella thermoacetica	1525	Complete Genome	2451
Rhodospirillum rubrum	1085	Complete Genome	3835
Salinibacter ruber M31	146919	Complete Genome	2812
Streptomyces coelicolor	1902	Complete Genome	8032
Synechocystis	1142	Complete Genome	3507
Thermanaerovibrio acidaminovorans	525903	Complete Genome	1737
Thermodesulfovibrio yellowstonii	28262	Complete Genome	1982
Thermotoga maritima	2336	Complete Genome	1852
Thermus thermophilus	274	Complete Genome	2227
Archaea			

Bathyarchaeota archaeon BA2	1700837	Scaffold	2426
Candidatus Micrarchaeum acidiphilum	425595	Contig	1761
Halobacterium salinarum NRC-1	2242	Complete Genome	1029
Haloterrigena turkmenica	543526	Complete Genome	5113
Korarchaeum cryptofilum OPF8	498846	Complete Genome	1602
Lokiarchaeum sp GC14	1655637	Contig	5378
Methanocaldococcus jannaschii	2190	Complete Genome	1787
Methanococcus maripaludis S2	39152	Complete Genome	1722
Methanopyrus kandleri AV19	190192	Complete Genome	1687
Methanothermobacter thermautotrophicus	145262	Complete Genome	1868
Nanoarchaeum equitans Kin4-M	160232	Complete Genome	536
Nanosalina sp	889948	Scaffold	1673
Natronobacterium gregoryi SP2	44930	Contig	3624
Nitrososphaera gargensis Ga9.2	1237085	Complete Genome	3523
Pyrobaculum aerophilum	13773	Complete Genome	2590
Thermococcus kodakarensis	311400	Complete Genome	2301
Eukarya			
Arabidopsis thaliana	3702	Complete Genome	31477
Caenorhabditis elegans	6239	Complete Genome	26596
Chlamydomonas reinhardtii	3055	Scaffold	14337
Dictyostelium discoideum	44689	Complete Genome	12746
Drosophila melanogaster	7227	Complete Genome	22005
Tetrahymena thermophila SB210	5911	Scaffold	26976
Amphimedon queenslandica	400682	Scaffold	29758
Ashbya gossypii	33169	Complete Genome	4760
Batrachochytrium dendrobatidis	109871	Complete Genome	8610
Capsaspora owczarzaki	595528	Scaffold	9794
Cyanidioschyzon merolae	45157	Complete Genome	4995
Emiliania huxleyi	2903	Scaffold	35697
Encephalitozoon cuniculi	6035	Complete Genome	2008
Entamoeba histolytica	5759	Scaffold	7959
Guillardia theta	55529	Complete Genome	24590
Homo sapiens	9606	Complete Genome	70225
Leishmania braziliensis	5660	Complete Genome	8084
Monosiga brevicollis	81824	Scaffold	9188
Phaeodactylum tricornutum	556484	Complete Genome	10465
Phytophthora ramorum	164328	Scaffold	15349
Plasmodiophora brassicae	37360	Complete Genome	9720
Plasmodium falciparum 3D7	5833	Complete Genome	5353

Schistosoma mansoni	6183	Complete Genome	11723
Thelohanellus kitauei	669202	Scaffold	14792
Trichoplax adhaerens	10228	Scaffold	11520
Ustilago maydis	5270	Contigs	6806

Table S5. Summary of disorder and TMH statistics for PVCs and non-PVCs.

STATISTICS	Disorder content (%)	LDR content (%)	HD proteins (%)	TM proteins (%)
	Mean/Median	Mean/Median	Mean/Median	Mean/Median
Planctomycetes	12.15/12.39	2.83/2.915	5.51/5.14	23.73/24.005
Verrucomicrobia	7.43/7.08	1.27/1.09	2.82/2.67	24.73/23.55
Chlamydiae	4.95/5.135	1.41/1.54	2.24/2.14	24.44/24.43
Lentisphaera	5.39/5.39	0.70/0.7	1.74/1.74	20.89/20.89
non-PVC Bacteria	6.82/5.945	1.44/1.0	2.52/1.85	24.08/23.885
Archaea	6.33/2.86	1.29/0.5	3.05/1.12	20.34/21.06
Eukaryote	20.50/20.4	10.92/10.11	11.90/10.85	20.61/19.26

P-values of Disorder Content comparisons by Wilcoxon Rank Sum Test (2-tailed)							
	Planctomycetes	**Verrucomicrobia**	**Chlamydia**	**Lentisphaera**	**non-PVC Bacteria**	**Archaea**	**Eukaryote**
Planctomycetes	-	7.49E-004	6.64E-004	2.01E-001	1.04E-004	3.05E-003	6.71E-005
Verrucomicrobia	7.49E-004	-	1.12E-002	3.11E-001	1.64E-001	3.97E-002	9.90E-006
Chlamydiae	6.64E-004	1.12E-002	-	1.14E-001	8.95E-002	1.54E-001	4.40E-006
Lentisphaera	2.01E-001	3.11E-001	1.14E-001	-	6.74E-001	5.35E-001	9.51E-002
non-PVC Bacteria	1.04E-004	1.64E-001	8.95E-002	6.74E-001	-	1.15E-001	5.48E-009
Archaea	3.05E-003	3.97E-002	1.54E-001	5.35E-001	1.15E-001	-	5.08E-005
Eukaryote	6.71E-005	9.90E-006	4.40E-006	9.51E-002	5.48E-009	5.08E-005	-
Threshold: 0.05	5.00E-002						

Table S5 (continued). Summary of disorder and TMH statistics for PVCs and non-PVCs.

P-values of LDR Content comparisons by Wilcoxon Rank Sum Test (2-tailed)							
	Planctomycetes	**Verrucomicrobia**	**Chlamydia**	**Lentisphaera**	**non-PVC Bacteria**	**Archaea**	**Eukaryote**
Planctomycetes	-	8.60E-006	1.59E-005	1.00E-001	5.85E-009	2.82E-006	2.33E-008
Verrucomicrobia	8.60E-006	-	1.08E-004	1.11E-001	9.63E-007	3.44E-005	2.02E-006
Chlamydiaee	1.59E-005	1.08E-004	-	1.14E-001	2.30E-006	5.55E-005	4.40E-006
Lentisphaera	1.00E-001	1.11E-001	1.14E-001	-	9.29E-002	1.07E-001	9.51E-002
non-PVC Bacteria	5.85E-009	9.63E-007	2.30E-006	9.29E-002	-	1.90E-007	7.81E-011
Archaea	2.82E-006	3.44E-005	5.55E-005	1.07E-001	1.90E-007	-	4.78E-007
Eukaryote	6.86E-008	2.02E-006	7.28E-006	1.58E-001	1.97E-010	1.04E-005	-
Threshold: 0.05	5.00E-002						

P-values of HD Protein Content comparisons by Wilcoxon Rank Sum Test (2-tailed)							
	Planctomycetes	**Verrucomicrobia**	**Chlamydia**	**Lentisphaera**	**non-PVC Bacteria**	**Archaea**	**Eukaryote**
Planctomycetes	-	8.60E-006	1.59E-005	1.00E-001	5.85E-009	2.82E-006	2.33E-008
Verrucomicrobia	8.60E-006	-	1.08E-004	1.11E-001	9.63E-007	3.44E-005	2.02E-006
Chlamydiae	1.59E-005	1.08E-004	-	1.14E-001	2.30E-006	5.55E-005	4.40E-006
Lentisphaera	1.00E-001	1.11E-001	1.14E-001	-	9.29E-002	1.07E-001	9.51E-002
non-PVC Bacteria	5.85E-009	9.63E-007	2.30E-006	9.29E-002	-	1.90E-007	7.81E-011
Archaea	2.82E-006	3.44E-005	5.55E-005	1.07E-001	1.90E-007	-	1.20E-006
Eukarayote	1.24E-007	3.29E-006	8.58E-006	1.58E-001	2.67E-010	4.47E-005	-
Threshold: 0.05	5.00E-002						

Table S5 (continued). Summary of disorder and TMH statistics for PVCs and non-PVCs.

P-values of TM Protein Content comparisons by Wilcoxon Rank Sum Test (2-tailed)							
	Planctomycetes	**Verrucomicrobia**	**Chlamydia**	**Lentisphaera**	**non-PVC Bacteria**	**Archaea**	**Eukaryote**
Planctomycetes	-	4.18E-001	3.88E-001	2.01E-001	8.01E-001	3.47E-003	6.05E-003
Verrucomicrobia	4.18E-001	-	9.44E-001	1.11E-001	6.66E-001	1.06E-003	1.16E-002
Chlamydiae	3.88E-001	9.44E-001	-	1.14E-001	7.23E-001	3.56E-003	1.48E-002
Lentisphaera	2.01E-001	1.11E-001	1.14E-001	-	2.08E-001	9.01E-001	7.97E-001
non-PVC Bacteria	8.01E-001	6.66E-001	7.23E-001	2.08E-001	-	5.00E-004	6.92E-004
Archaea	3.47E-003	1.06E-003	3.56E-003	9.01E-001	5.00E-004	-	7.66E-001
Eukaryote	6.05E-003	1.16E-002	1.48E-002	7.97E-001	6.92E-004	7.66E-001	-
Threshold: 0.05	5.00E-002						

Table S5 (continued). Summary of disorder and TMH statistics for PVCs and non-PVCs.

Still functionally uncharacterized

The three methods used here failed to provide any information for 44,219 proteins (25.5% of the whole dataset). There are various explanations for this. In some cases, they might be dubious proteins derived from wrong predictions during the automated genome analysis. While in other cases, they could represent very interesting proteins that we currently know nothing about.

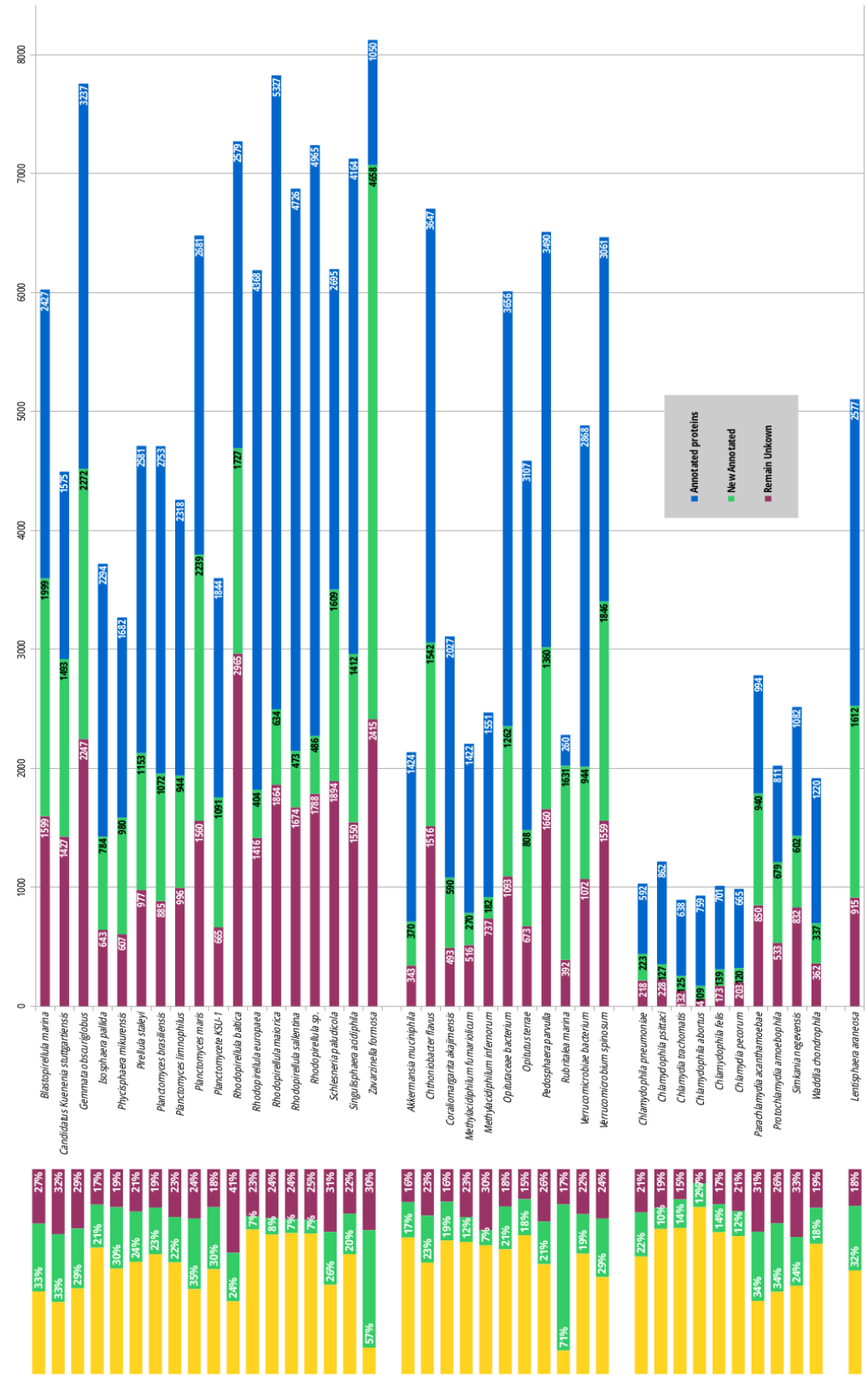


Figure 13. Number of proteins with improved annotations

4.2

Planctomycetes attached to algal surfaces: insight into their genomes

Genomics 110 (2018) 231–238

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Planctomycetes attached to algal surfaces: Insight into their genomes

Mafalda Faria[a,1], Nicola Bordin[b,1], Jana Kizina[c], Jens Harder[c], Damien Devos[b], Olga M. Lage[a,d,*]

[a] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal
[b] Centro Andaluz de Biología del Desarollo, CSIC, Junta de Andalucía, Universidad Pablo de Olavide, Carretera de Utrera, Km. 1, 41013 Seville, Spain
[c] Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany
[d] CIMAR/CIIMAR – Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n, 4450-208 Matosinhos, Portugal

ARTICLE INFO

Keywords:
Genome
Roseimaritima ulvae
Rubripirellula obstinata
Mariniblastus fucicola
Lifestyle in macroalgal biofilm
Huge proteins

ABSTRACT

Planctomycetes are bacteria with complex molecular and cellular biology. They have large genomes, some over 7 Mb, and complex life cycles that include motile cells and sessile cells. Some live on the complex biofilm of macroalgae. Factors governing their life in this environment were investigated at the genomic level. We analyzed the genomes of three planctomycetes isolated from algal surfaces. The genomes were 6.6 Mbp to 8.1 Mbp large. Genes for outer-membrane proteins, peptidoglycan and lipopolysaccharide biosynthesis were present. *Rubripirellula obstinata* LF1^T, *Roseimaritima ulvae* UC8^T and *Mariniblastus fucicola* FC18^T shared with *Rhodopirellula baltica* and *R. rubra* SWK7 unique proteins related to metal binding systems, phosphate metabolism, chemotaxis, and stress response. These functions may contribute to their ecological success in such a complex environment. Exceptionally huge proteins (6000 to 10,000 amino-acids) with extracellular, periplasmic or membrane-associated locations were found which may be involved in biofilm formation or cell adhesion.

1. Introduction

Planctomycetes are bacteria belonging to the *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* (PVC) superphylum [1] that are found in a myriad of ecosystems which highlights their environmental relevance. Their presence has been detected, usually in low amounts, in common aquatic and terrestrial habitats but also in extreme environments and in association with other organisms ([2] and references therein). Due to a great metabolic diversity, planctomycetes are considered to play an important role in global environmental cycles, contributing to the global carbon [3], nitrogen and sulfur cycles. The striking phenotypical traits and cell biology of many members of *Planctomycetes* are remarkable. For example, some of them divide by budding [4] despite the fact that the cell division protein FtsZ (Filamenting temperature-sensitive mutant Z), otherwise present in the vast majority of bacteria, is not found in any planctomycetal proteome [5]. They also have a complex life cycle and comprise a complex cell plan uncommon in bacteria due to extensive endomembrane development [6–8]. Furthermore, endocytosis and the presence of membrane coat-like proteins have been described in these organisms [9,10]. Although they were considered to be peptidoglycan-less bacteria for many years [11] recent structural and genetic evidence gave support to a different concept [12,13] and in 2015, peptidoglycan was observed in the cell wall of five planctomycetes strains [14,15]. Due to their peculiar cell biology, planctomycetes have been a topic of interest since early in the genomic era. *Rhodopirellula baltica* SH1^T was the first sequenced genome belonging to the *Planctomycetes* [3]. Since then, the sequencing of other *Planctomycetes* genomes has opened a huge range of possibilities to better understand these bacteria.

Rubripirellula obstinata LF1^T, *Roseimaritima ulvae* UC8^T, and *Mariniblastus fucicola* FC18^T were isolated from the macroalgal biofilm of *Laminaria* sp., *Ulva* sp. and *Fucus spiralis* respectively, sampled from the north coast of Portugal [16]. Strains FC18^T, LF1^T and UC8^T have been taxonomically characterized as a new genera of *Planctomycetes* [17,18]. Strains LF1^T and FC18^T were the only isolates obtained of their genera in the isolation experiments while three isolates were retrieved from *R. ulvae*. Strains closely related to UC8^T were found in the biofilm on seawater reverse osmosis membranes (GenBank: HQ326270) and in the sponge *Niphates* sp. [19] of Moreton Bay, Australia. Furthermore, strains of *R. ulvae* have been found associated with other macroalgae, namely *Porphyra dioica* [20] and *Chondrus crispus* [21]. *R. obstinata* was also found associated with other macroalgae: *Mastocarpus stelatus* [20], *Ulva* sp., *Chondrus crispus* and *Fucus spiralis* [19] but no close relatives were found in other environments. *M. fucicola* was frequently detected

* Corresponding author at: Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal.
E-mail address: olga.lage@fc.up.pt (O.M. Lage).
[1] The authors contributed equally.

http://dx.doi.org/10.1016/j.ygeno.2017.10.007
Received 24 August 2017; Received in revised form 6 October 2017; Accepted 21 October 2017
Available online 23 October 2017

associated with the macroalgae *Ulva* sp., *Chondrus crispus*, *Sargassum muticum* and *Porphyra dioica* [20–21]. Moreover, related strains with a 98% 16S rRNA gene identity were isolated from the phycosphere of *Enteromorpha prolifera* in the Qingdao Sea (GenBank: JF769591 and JF769639). In this study, we present the draft genomes of *Rubripirellula obstinata* LF1^T, *Roseimaritima ulvae* UC8^T and *Mariniblastus fucicola* FC18^T. Special relevance was given to particular characteristics, including ones related to their lifestyle in algal biofilms.

2. Methods

2.1. Biological material

Roseimaritima ulvae UC8^T (GenBank: HQ845508.1), *Rubripirellula obstinata* LF1^T (GenBank: DQ986201.2) and *Mariniblastus fucicola* FC18^T (GenBank: HQ845450.1), were isolated from the macroalgal biofilm of *Ulva* sp. sampled in Carreço (41°44′N, 8°52′W), *Laminaria* sp. in Porto (41°19′N, 8°40′W) and *Fucus spiralis* from Carreço (41°44′N, 8°52′W), respectively [16–18].

2.2. Genomic DNA extraction and 16S rRNA gene amplification and analysis

Genomic DNA was obtained from batch cultures on modified solid M13 [16] at 24 °C and extracted in duplicate using the E.Z.N.A. ® Genomic DNA Isolation Kit (Omega Bio−Tek, VWR). The 16S rRNA gene was amplified using 1 μl of the extracted gDNA, cooled on ice with 2 μM of the universal primers 27F and 1492r [22] in 25 μl of a polymerase chain reaction (PCR) mixture (1 × PCR buffer, 1.5 mM $MgCl_2$, 1 unit of GoTaq Flexi DNA Polymerase (Promega), 200 μM of each deoxynucleoside triphosphate (dNTPs)). The PCR program was performed in a MyCycler™ Thermo Cycler (Bio-Rad) and consisted in an initial denaturing step of 5 min at 95 °C; 30 cycles of 1 min at 94 °C; 1 min at 52 °C; 90 s at 72 °C; and a final extension of 5 min at 72 °C. PCR products (5 μl) were visualized after electrophoresis in a 1.2% agarose gel in 1 × Tris base, boric acid and EDTA - TBE buffer. The PCR products were purified and sequenced by Macrogen (Amsterdam) to confirm strains identity.

2.3. Next generation sequencing

Genomic DNA sequencing was performed using Illumina MiSeq technology by the Max Planck-Genome-centre in Cologne, Germany (http://mpgc.mpipz.mpg.de/home/). The genomic library preparation was performed with the NEB NextUltra™ DNA Library Prep Kit for Illumina, NEB. The Illumina method was performed in two (FC18^T and LF1^T) or three (UC8^T) runs, generating 250 bp long paired-end reads, obtaining 6,697,558, 6,856,066 and 6,437,529 reads respectively.

2.4. Raw data assembly

Raw reads were trimmed with SolexaQA v.2.2 [23] and DynamicTrim (trimming value of 10). After the trimming step, the paired-end reads were normalized using Khmer 1.0 [24] and assembled with VelvetOptimiser v 2.2.5 [25] and SPAdes v 3.1.0 [26]. For the assembly using IDBA-UD v 1.1.0 [27] the steps were very similar to the ones mentioned above, with an initial trimming step.

The contigs obtained from the first assembly of each strain were merged and *de novo* assembled in Sequencher v 4.6 (Gene Codes Corporation, Ann Harbor, USA). GENEious R8 (Biomatters, Auckland, New Zealand) [28] was also used to identify possible contig elongations. The contigs used in the mapping were the two or three longest contigs obtained in Sequencher 4.6 output.

2.5. Automatic annotation

Annotation of the contigs was performed using the Prokka v1.11 pipeline [29] with default parameters. Prokka consists of several general-purpose tools such as Basic Local Alignment Search Tool - BLAST + and HMMER, as well as tools specifically tailored for prokaryotic tRNAs and opening reading frames, such as Aragorn and Prodigal.

2.6. Homologous protein clusters – differential analysis

Protein clusters present in all strains were identified using the OrthoMCL v2.0 suite [30]. We followed the suggested protocol by setting the all-*vs*-all BLAST *E*-value threshold at $1e^{-5}$.

The proteomes of strains FC18^T, UC8^T and LF1^T were differentially compared to the ones of *Blastopirellula marina*, *Planctopirus limnophila* and *Zavarzinella formosa*. Afterwards, we detected the proteins in the clusters shared between each bacterium or was unique to a specific strain. The OrthoMCL output was parsed using three Python scripts: one for obtaining all the clusters in common in a 6 bacteria (LF1^T, UC8^T, FC18^T, *R. baltica*, *B. marina* and *P. limnophila*) all-*vs*-all comparison, one for obtaining all the clusters in a 3-*vs*-3 (LF1^T, UC8^T, FC18^T) all-*vs*-all comparison, and a last script that, provided with the results from the previous two scripts, extracted the FASTA sequences for each of the clustered proteins.

In order to identify proteins specifically related to the macroalgal biofilm environment another differential analysis was performed, using the same parameters as in the first differential analysis, between planctomycetal strains associated with macroalgae (*R. baltica*, *Phycisphaera mikurensis*, *R. rubra* SWK7, strain FC18^T, strain LF1^T and strain UC8^T) and planctomycetes from other environments, *B. marina*, *P. limnophila*, *Zarvazinella formosa* and *Singulisphaera acidiphila*.

2.7. Protein identification

After determining the common proteins among strains LF1^T, UC8^T and FC18^T, the ones that resulted from the two reference genomes (*R. baltica* and *P. limnophila*) were identified using UniProt (http://www.uniprot.org) and GenBank (*B.marina*) (http://www.ncbi.nlm.nih.gov/genbank/) descriptions. For the identification of the proteins shared between FC18^T, LF1^T and UC8^T (not belonging to any database) InterProScan [31], PSI-BLAST, and the initial Prokka annotation were used. The results were manually curated. The shared genes were assessed using InterPro v 5.14–53.0 and a pipeline developed in-house. Gene ontology terms (GO terms) were assigned to the proteins in common among strains LF1^T, UC8^T and FC18^T using Blast2GO Basic v 3.1.3 [32].

2.8. Subcellular location of proteins

Putative subcellular localization of the proteins with > 6000 a.a. was determined using PSORTb (version 3.0.3) by selecting prediction specific for Gram-negative bacteria [33]. Furthermore, these proteins were tested for the presence of membrane spanning domains using TMHMM Server v.2.0 [34].

2.9. Contigs realignment

CONTIGuator 2 [35] was used to realign the contigs. This tool is based on mapping the contigs against a reference genome – *R. baltica* and *B. marina* in this case, using blastn with an *E*-value of 1e-5. In order to confirm the results obtained from CONTIGuator, several approaches were performed to assess the accuracy and validity of the result including ABACAS [36], Mauve [37], MUMmer and PROmer [38].

2.10. Prophage sequences detection

The detection of prophage sequences within the three strains was performed with PHAST [39].

2.11. Phylogenetic assignment (RpoB protein)

Phylogenetic profiling of the three strains was performed using the RNA-polymerase subunit beta protein, encoded by *rpoB*. Protein sequences were extracted using BLAST [40] with an E-value of 1e-5. The multiple sequence alignment was created using ClustalOmega [41] with default parameters and manually curated in Jalview [42]. The tree was generated using PhyML 3.1 [43] using the LG matrix, 100 bootstraps, tree and leaves refinement, SPR moves, and amino acids substitution rates determined empirically.

2.12. Search for proteins related to outer membrane biomarkers, cytochrome and V-type ATPases

Evidence for these proteins was obtained running PSIBLAST on the proteomes of the 3 strains for 3 iterations with an E-value cutoff of $1e^{-5}$.

2.13. Scanning electron microscopy

Cells of strain FC18^T were fixed in 2.5% glutaraldehyde in 0.2 M cacodylate buffer and in 2% osmium tetroxide in the same buffer, serially dehydrated in ethanol and air-dried before observation in a JEOL JSM 6301F scanning electron microscope.

3. Results and discussion

3.1. Phylogeny

The phylogenetic relationship of the three strains based on the analysis of the 16S rRNA gene [16] was reinvestigated based on the analysis of the beta subunit of the RNA polymerase (*rpoB*) which has been suggested as a novel molecular marker to infer phylogeny in *Planctomycetales* [44]. *Rubripirellula obstinata* LF1^T, *Roseimaritima ulvae* UC8^T and *Mariniblastus fucicola* FC18^T are phylogenetically related to *Rhodopirellula baltica* SH1^T, with strain LF1^T being most closely related, followed by strains UC8^T and FC18^T (Fig. 1). ANI values of < 69% among the three strains and closest relatives show a large genetic distance between the strains.

3.2. General overview of the genomes

The genome sizes of *Rubripirellula obstinata* LF1^T, *Roseimaritima ulvae* UC8^T and *Mariniblastus fucicola* FC18^T vary between 6.6 Mbp and 8.1 Mbp (Table 1), within the range of previous observations for genomes belonging to the *Planctomycetaceae* [45]. The software CheckM [46] showed a very low level of contamination of the genomes (1.16% for LF1^T, 0% for UC8^T and 0.11% for FC18^T), which are typical values for very good draft genomes. Within the *Planctomycetaceae* the G + C content varies from 50 to 67% [45], the values obtained for the three strains: 54.1% for LF1^T, 59.12% for UC8^T and 53.40% for FC18^T, are in this range. These *in silico* values confirmed previously determined G + C content of the three strains [17,18]. In the genomes of the three planctomycetes, 5913, 5943 and 5894 open reading frames (ORFs) were identified for LF1^T, UC8^T and FC18^T, respectively (Table 1). In comparison, *R. baltica* SH1^T has 7325 putative protein encoding ORFs [41] and *B. marina* DSM 3645^T has 6025 [42].

The detection of orthologous and paralogous proteins conserved among strains LF1^T, UC8^T and FC18^T was performed after annotation and gene prediction with Prokka [29]. The three strains have a total of 6187 proteins classified in common clusters. The number of proteins unique to each strain is 1510 for LF1^T, 1925 for UC8^T and 1932 for FC18^T and the number of paralogues is, respectively, 516, 385 and 290 (Additional file 1: Table S1). Strains FC18^T and LF1^T share the lowest number of clustered proteins, approximately 280, followed by FC18^T and UC8^T with 541 and 562. Strains UC8^T and LF1^T are the ones that share more clustered coding DNA sequences (CDSs), 811 and 832 respectively. These results support the *rpoB* gene phylogenetic closeness of UC8^T and LF1^T and a larger distance between these two and FC18^T (Fig. 1). In relation to the orthologous proteins shared by the three strains, the values are quite similar (Additional file 1: Table S1).

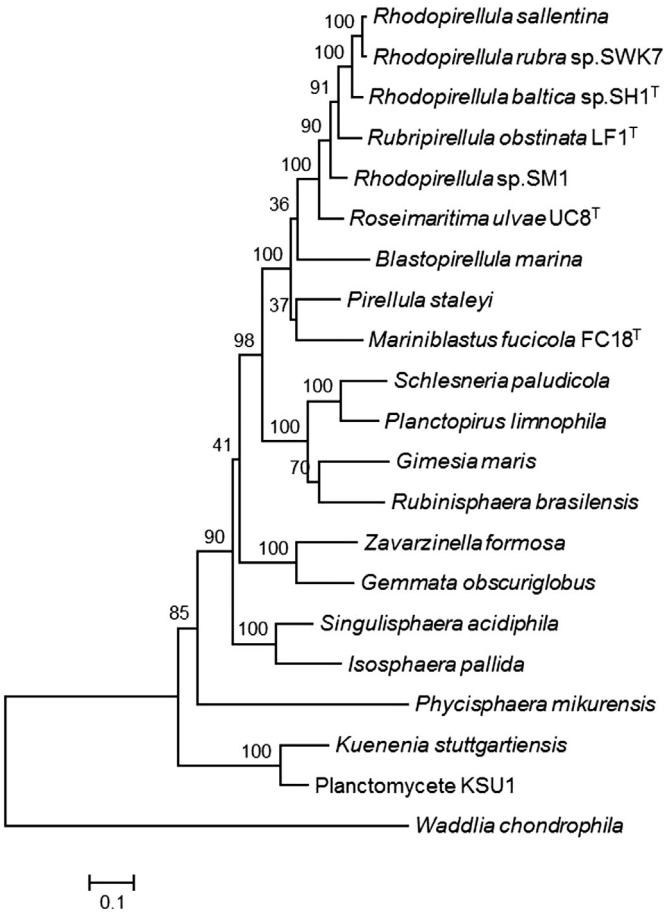


Fig. 1. Phylogenetic profiling of the 3 sequenced strains and their relationship with *Planctomycetes* based on the analysis of the *rpoB* gene. *Waddlia chondrophila* (*Chlamydia*) was used as outgroup for tree rooting. Bootstrap values are shown at each branch node and divergence can be calculated using the branch length provided.

Table 1
General overview of the genome features from strains FC18^T, LF1^T and UC8^T.

Attribute	Strains		
	LF1^T	UC8^T	FC18^T
Genome size (bp)	6,588,559	8,130,296	6,539,195
Contamination	1.16%	0.00%	0.11%
DNA G + C content (%)	54.1	59.12	53.4
CDS - Prokka (bp)	3958	4479	3543
tRNA genes	69	71	66
Contigs	309	108	64
ORFs	5200	5759	5096

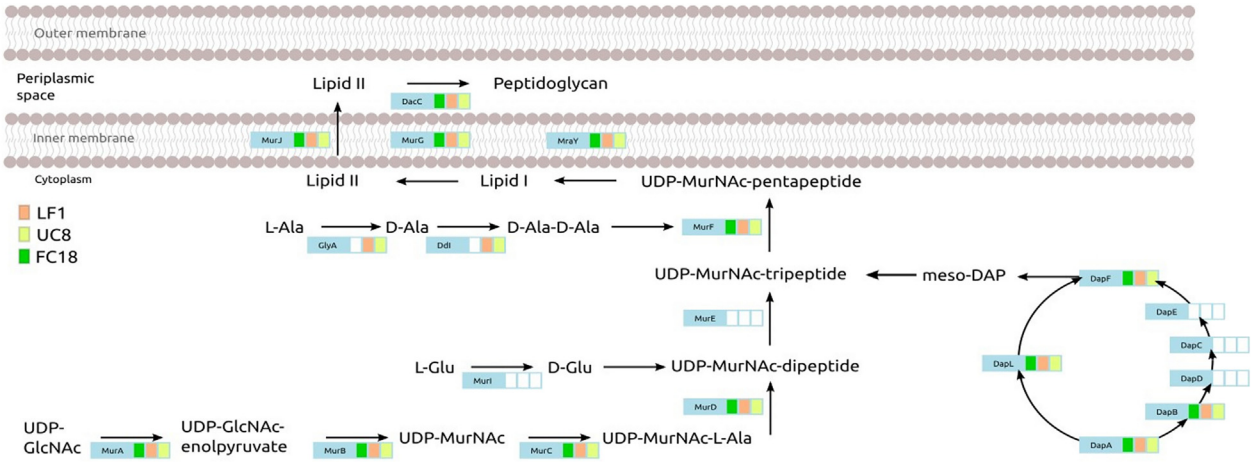


Fig. 2. Peptidoglycan biosynthesis pathway. Evidence for genes encoding enzymes for the biosynthesis of peptidoglycan. At the side of each gene involved in this process a square is filled with different colors if present (orange for LF1^T, yellow for UC8^T and green for FC18^T) or blank if absent or not detected through sequence similarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Genetic evidences of a diderm (Gram-negative) cell wall in strains FC18^T, LF1^T and UC8^T

The *in silico* proteomes of the three strains were searched for the presence of outer membrane biomarkers as referred by Speth et al. [13]. Peptidoglycan synthesis related proteins were searched against the ones of *Planctopirus limnophila,* previously named *Planctomyces limnophilus* (Fig. 2 and Additional file 1: Table S2). Relatives of the peptidoglycan precursor synthesis proteins MurA, MurB, MurC, MurD, MurF, MurG, MurJ, MraY and of peptidoglycan synthesis proteins DapA, DapB, DapF, DapL were present in the 3 strains. FC18^T and UC8^T have the enzyme glucosamine-1-phosphate *N*-acetyltransferase (EC 2.3.1.157) and LF1^T has UDP-*N*-acetylmuramoylalanyl-D-glutamate-2,6-diaminopimelate ligase (EC 6.3.2.13) both involved in peptidoglycan biosynthesis.

Furthermore, the cell elongation protein MreB, the outer-membrane invagination protein TolQ, and the cell division protein FtsK were also present. Regarding the outer membrane biomarkers, proteins related to LPS insertion complex and outer membrane protein (OMP) insertion were detected in the three genomes. Furthermore, a TonB system, which is bacterial outer membrane proteins binding and transporting ferric chelates, vitamin B (12), nickel complexes, and carbohydrates, was only identified in strain UC8^T.

Our results are consistent with previous studies that demonstrated by *in silico* [13] and experimental [14,15] analyses that planctomycetes possess peptidoglycan and an outer membrane typical of a Gram-negative cell wall.

3.4. Clustered regularly interspaced short palindromic repeats (CRISPR)

CRISPR regions are essential in the adaptive immunity of some bacteria and archaea by responding to and eliminating invading genetic material like bacteriophages and conjugative plasmids [47]. The genomes of strains FC18^T and LF1^T have a cluster of CRISPR genes that are separated from the CRISPR-associated endonuclease Cas9 (Additional file 1: Table S3). We could not detect CRISPR-related genes in strain UC8^T. However, the three strains possess in their genomes several proteins related to phage elements like a phage major capsid protein (data not shown).

3.5. Vacuolar-type ATPases (V-type ATPase)

ATPases are membrane-associated machines that couple the transfer of protons or sodium cations across the membrane with ATP hydrolysis or synthesis. In bacteria, ATPases comes in two flavors, V- or F-type [48]. The three strains have ATPases. However, V-type ATPase related proteins are only found in strains UC8^T and LF1^T. The seven proteins found in each strain are organized in a cluster/operon. (Additional file 1: Table S4). Curiously, strain FC18^T only shares a similar F-type ATPase protein composition with strains UC8^T and LF1^T. V-type ATPases are highly conserved evolutionary membrane-bound rotary motor proteins commonly found in eukaryotic cells and in some bacteria like members of *Firmicutes, Fusobacteria, Spirochaetes, Chlamydiales, Thermus/Deinococcus* and *Thermotogae* [49]. Further sequences related to V-type ATPase were also found in member of *Proteobacteria, Bacteroidetes, Actinobacteria, Cyanobacteria, Mycoplasma, Planctomycetes,* and *Verrucomicrobia* as obtained from a GenBank search. In *Planctomycetales*, *R. europaea* SH398, *R. europaea* 6C, *R. rubra* SWK7, *R. sallentina* SM41, *Rubinisphaera*, *Gemmata*, *Pirellula*, *Zavarzinella*, *Isosphaeraceae* and *Candidatus* Brocadiaceae (Jettenia, Kuenenia, Scalindua and Brocadia) also possess V-type ATPases clusters of various proteins.

3.6. Cytochrome related proteins

The number of cytochrome related proteins in the three genomes vary greatly. (Additional file 1: Table S5). Strain FC18^T possesses 47, strain LF1^T 27 and strain UC8^T 103. This phenomenon is present also in other planctomycetes, with numbers ranging from 13 in *Gemmata obscuriglobus* to 78 in *Rhodopirellula rubra* SWK7. The majority of these proteins are annotated as “cytochrome C related”, but strain UC8^T also contains cytochrome *bd*-I and bd-II ubiquinol oxidase subunits and a succinate dehydrogenase cytochrome *b*558 subunit. Strain FC18^T additionally contains a biotin biosynthesis cytochrome P450. In *Escherichia coli*, Cytochrome bd-I ubiquinol oxidase subunit 1 is related to the production of a proton motive force in the inner membrane and is the predominant aerobic respiratory chain under low aeration growing conditions. The succinate dehydrogenase cytochrome *b*558 subunit protein is involved in the tricarboxylic acid cycle in *Bacillus subtilis*. A comparable number to the one obtained in strain UC8^T genome was observed in *Geobacter uraniireducens* (104 cytochromes) [50]. However the rationale for such a high cytochrome number is still to be investigated.

3.7. Uniqueness of macroalgae-associated planctomycetes

Aiming to find a proteome pattern common to planctomycetes associated to macroalgae (strains FC18^T, LF1^T, UC8^T, *R. rubra* SWK7 and *R. baltica.* The latter, even though isolated from the water columns, is also associated with macroalgae [16]), a comparative *in silico* analysis of this group was performed against *Blastopirellula marina*, *Planctopirus*

limnophila and *Zavarzinella formosa* proteomes, which are planctomycetes isolated from brackish water in the Baltic Sea [51], from the freshwater lake Plußsee in Holstein [52] and an acidic *Sphagnum* peat bog [53] respectively. One hundred and fifty seven proteins are shared among the five species (Additional file 1: Table S6). Out of these, 45 (28.7%) were hypothetical (of these, 15 were unrelated to any putative function). Thirty seven (23.6%) could be related to cell membrane or outer membrane trafficking. Other functions detected were related to carbohydrate metabolism, lipid metabolism, nucleic acid metabolism, cell signaling, adhesion, stress response, LPS biogenesis, metal resistance (tellurium), cell homeostasis, response to ethylene, oxygen binding, and nitrogen metabolism. Macroalgae, like plants, also produce ethylene [54,55] and the macroalgal associated planctomycetes possess proteins related to ethylene production (Additional file 1: Table S6). The clustering between planctomycetal strains associated to macroalgae was expanded to include more divergent species (*P. mikurensis*, *R. rubra* SWK7, *R. baltica*, strain FC18^T, strain LF1^T and strain UC8^T) and planctomycetes from other environments (*B. marina*, *P. limnophila, Zarvazinella formosa* and *Singulisphaera acidiphila*). This showed 28 protein clusters that are unique to the macroalgae environment, with functions related to membrane permeability or attachment, subunits of channels for importing/exporting ions, or mucin-related proteins (Additional file 1: Table S7). All of them potentially help to maintain the biofilm or are involved in the symbiosis with the algae.

3.8. Sulfatases and other polysaccharide-degrading enzymes

A high number of sulfatase genes have been found in planctomycetes since the annotation of their first genome [3]. Marine macroalgae are massive producers of sulfated polysaccharides like fucoidans in brown algae, carrageenans in red algae, and ulvans in green algae [56]. As sulfatases have the potential to hydrolyze sulfate esters and sulfamates they play an important role in the sulfur cycle in marine environments. Moreover, marine planctomycetes, especially if in association with macroalgae, may nutritionally benefit from these polysaccharides. In the genomes of strains FC18^T, LF1^T and UC8^T, 61, 36 and 95 sulfatase proteins were detected, respectively (Additional file 1: Table S5) using as a query the reference sequences referred by Wegner et al. [57]. All *Rhodopirellula* related strains showed a number of sulfatase encoding genes higher than 100. *R. rubra* strain SWK7, a strain isolated from macroalgae surface, exhibited 196 sulfatases. However, with the exception of strain UC8^T that possesses 98 sulfatases, the other two strains showed lower numbers (61 in FC18^T and 36 in LF1^T) which are closer to the ones observed in *B. marina* (40) and *P. staleyi* (34). The number of sulfatase encoding genes (SEGs) in other planctomycetes species varied between 12 in *Gemmata* and 83 in *Planctomyces maris*. With the exception of *Blastopirellula marina* that only has 40 SEGs, all the species with a high number of sulfatases are marine.

Strain LF1^T possesses cellulase, amylase, agarase, xylanase, porphyranase, and *N*-acetylglucosamine deacetylase (Additional file 1: Table S5). Comparatively, UC8^T lacks xylanase and FC18^T agarase. Strain FC18^T has a high number of xylanases. No proteins related to the degradation of pectin, lignin, ulvan, carrageenan, fucoidan or laminarin were observed. λ-carrageenase was not found in any of the three genomes but a pre-lambda-carrageenase protein exists in the genome of *Rhodopirellula rubra* SWK7. Strain LF1^T encodes for a broad range of enzymes for polysaccharide utilization but lower number of sulfatases. Curiously UC8^T isolated from *Ulva* sp. cannot utilize ulvan, a polysaccharide produced by this alga. The three sequenced strains are, however, well equipped to utilize several of the polysaccharides produced by macroalgae.

3.9. Stress responses

Metal scavenging proteins are well represented in the genomes of the three strains (Additional file 1: Table S5). These include copper related proteins, Co-Zn-Cd resistance proteins, Cd transporting ATPase, mercury resistance proteins, arsenate related proteins, and a considerable number of multidrug resistance proteins. Furthermore, superoxide dismutase, catalase-peroxidase, a SOS-response protein, glutathione related proteins, several compatible osmolites (betaine, glycine, proline and trehalose related proteins), heat-shock proteins, and UV resistance proteins are also present in the three strains. The three strains are protected by the presence of D-tyrosyl-tRNA deacylase against a potential noxious effect of D-tyrosine. Strain UC8^T showed overall the highest number of stress related proteins. Common osmolytes also present in other bacteria include glycine betaine (*N,N,N*-trimethylglycine), proline, ectoine (1,4,5,6-tetrahydro-2-*methyl*-4-pyrimidine carboxylic acid), and trehalose [58]. Trehalose has the added advantage of being an antioxidant.

Living in the biofilm of macroalgae in rocky beach pools, these planctomycetes are subjected to stressful conditions. These include fluctuations in salinity due to tidal variation, high exposure to UV radiation, oscillations in temperature, and pollution due to anthropogenic input in coastal areas. Furthermore, as they inhabit biofilms, these cells have to cope with various complex interactions induced by the macroalgae or by other microorganisms, such as, escaping the action of oxygen radicals through superoxide dismutase, catalase-peroxidase and glutathione (Additional file 1: Table S5). Another important function for these bacteria is the capacity to control oxygen levels (by having oxygen binding proteins). The considerable high number of proteins somehow related to transport across the cell (Additional file 1: Table S6) can be expected due to the need of these planctomycetes to interact with the macroalgae, their presence in a biofilm, and their need to overcome environmental stress. Moreover, other proteins related to these aspects were also found, like cell signaling, adhesion, and stress response (Additional file 1: Tables S5 and S6).

In accordance with our results, Kim and collaborators [59] observed augmentation of genes related to stress responses in 3 genomes of planctomycetes inhabiting the blades of the macroalga *Porphyra umbilicalis*. Strain UC8^T is the one best prepared to cope with these environmental conditions since it possesses the highest number of stress response proteins.

3.10. Huge proteins in the three genomes

In the genomes of strains FC18^T, LF1^T and UC8^T several huge proteins with > 6000 amino acids are present (Additional file 1: Table S8). A comprehensive survey of 'giant genes' in planctomycetes has recently been published [60]. According to the prediction in TMHMM [34], none of these proteins possess transmembrane helices and may be considered secreted proteins. However, when we searched for the putative subcellular localization of these proteins using PSORTb on a scale of 0 to ten (Additional file 1: Table S8), only 4 proteins (the bifunctional hemolysin/adenylate cyclase precursors) were undoubtedly (10 points) considered as extracellular proteins. In 6 others, the probability to be extracellular was high (6.04 points). In another 3, periplasmic location was higher. Only one protein had a higher probability to be in the cytoplasm or in the outer membrane. Several of these proteins have tandem repetition of a domain (HemolysinCabind, Calx-beta or SdrD_B), which is indicative of a cytoplasmic membrane, cell wall, or extracellular location [61]. Variation in number of these domains is associated with the generation of antigenic and functional diversity among surface proteins [62–66].

Four of these proteins (LF1_00233 (8275 a.a.), LF1_01856 (6270 a.a.), UC8_01824 (8958 a.a.) and UC8_0358 (7047 a.a.)), were annotated as bifunctional hemolysin/adenylate cyclase precursors with a large number (24–32) of repeats of the hemolysinCabind domain. Proteins with this domain are known to be present in Gram-negative bacteria, are secreted into the growth medium, and are capable of binding calcium.

UC8_01824, the biggest protein found in the 3 genomes with almost 10,000 a.a. also has pentaxin, pentapeptide, lactonase, and PKD domains. Pentaxin proteins are involved in acute immunological responses [67] and are a class of pattern recognition receptors. The pentapeptide repeat proteins (PRP) were first identified in many cyanobacterial proteins but were also found in other bacteria and plant proteins. Their function is unknown [68]. Lactonases can be involved in the disruption of quorum sensing. PKD domains are found in extracellular parts of proteins like the archaeal surface layer proteins that protect the cell from extreme environments [69]. This domain was first identified in the polycystic kidney disease related polycystin-1 - PDK1 gene, which encodes for a large cell surface glycoprotein of unknown function [70]. It may be involved in protein–protein and protein–carbohydrate interactions.

The matrixin proteins LF1_03591 (6494 a.a.) and UC8_04175 (6003 a.a.) (Fig. 3) share a peptidase_M10 domain. Proteins with this domain are extracellular metalloproteases that cleave peptides, require zinc for catalysis, and degrade the extracellular matrix. According to PSORTb, they are 6.04 extracellular and 3.60 outer membrane (Additional file 1: Table S8). Homologues of these proteins are only found in the genus *Rhodopirellula*. Thus, we propose that these proteins are biological markers for this branch of the planctomycetes tree (Fig. 1) which also include *Rubripirellula* and *Roseimaritima*. The new genus of strain FC18^T, phylogenetically more distant, does not possess a corresponding protein, suggesting that *Mariniblastus fucicola* is not a member of this branch.

FC18_02694 (7782 a.a.) is a serine-aspartate repeat-containing protein D precursor with a series of 27 SdrD domain that should be involved in FC18^T attachment to the extracellular matrix and in the formation of biofilm. These functions were verified in *Staphylococcus aureus* [71]. Based on PSORTb, it is 6.04 extracellular and 3.60 outer membrane (Additional file 1: Table S8).

FC18_1343 (7202 a.a.) is a Calx-beta domain protein. This motif is present as a tandem repeat in the cytoplasmic domains of Calx sodium-calcium exchangers used to expel calcium from the cell and is not described as secreted proteins. It could be a periplasmic protein according to the highest probability found with PSORTb (6.86 periplamic and 3.01 extracellular; Additional file 1: Table S8).

Exoglucanase B proteins are known to hydrolyze cellohexaose. LF1_01536 (6783 a.a.), annotated as an exoglucanase B, has 10 DUF 5122 domains that are beta-propellers of unknown function, 3 fn3 domains, and a PKD domain. fn3 is a fibronectin type III, an evolutionary conserved protein domain, widely found in animal extracellular proteins but also in yeast, plants, and bacterial proteins.

UC8_04174 (6091 a.a.) has an unknown predicted function and a PSORTb location of 6.04 extracellular and 3.60 outer membrane (Additional file 1: Table S8). It contains three PF04151.13 PPC domains. These Plants and Prokaryotes Conserved (PPC) domains are found in bacteria, archaea, and plants. Previous studies indicate that this domain is essential for its nuclear location [72] but other functions remain unknown [73].

UC8_01761 (6601 a.a.) was annotated as a tRNA(Glu)-specific nuclease WapA precursor. It has a CARDB domain that stands for “Cell Adhesion Related Domain found in Bacteria” and six RHS (rearrangement hotspot)-repeat domains, which are conserved unique core sequence shared by large number of proteins. These proteins include secreted bacterial insecticidal toxins whose function is poorly understood. The Gram-negative Rhs proteins mediate intercellular competition by inhibiting the growth of neighboring cells [74]. Its location is 6.04 extracellular and 3.60 outer membrane (Additional file 1: Table S8).

FC18T_03125 annotated as a Laminin G domain protein has thirteen DUF5122 domains of beta-propellers of unknown function, three Laminin_G_3 domains, and three Cadherin_3 domains. It could be a periplasmic protein according to the highest probability found with PSORTb (6.86 periplamic and 3.01 extracellular). Laminins are extracellular matrix proteins and have an active role in cell differentiation, migration, and adhesion [75].

High molecular weight proteins like LapA protein of *Pseudomonas putida* (8682 amino acids) are a group of surface proteins important in the formation of biofilm that are generally designated Bap (biofilm-associated proteins) [76]. Reva and Tümmler [77] have also found giant bacterial genes in 47 taxa, including Planctomycetes. These were related to surface proteins or polyketide/non-ribosomal peptide synthetases, relevant to competition or adaptation to hostile environments. The findings of Kohn et al. [60] on the giant genes in the genome of the planctomycete *Fuerstia marisgermanicae* give further support to our results. These huge proteins may be related to the biofilm environment. A biofilm matrix is clearly present in these planctomycetes associated to macroalgae as can be seen for strain FC18^T (Fig. 4) and experimental evidence of the capacity of UC8^T to adhere to surfaces was obtained (data not shown).

4. Conclusion

The *in silico* analysis of the genomes of these three planctomycetes provide support for a Gram negative nature of the planctomycetes cell wall.

Our results on the genome analysis showed that strain UC8^T has the largest genome of the three stains, which is consistent with the higher protein number encountered for several functions, like stress conditions responses or cytochrome related proteins (103 proteins). On the contrary, strain LF1^T with a smaller genome showed a restricted number of proteins for many functions, such as cytochrome related proteins (only 27). This result is in agreement with the great absence of metabolic capability showed by this strain during the studies for its taxonomic characterization [17] and substantiates its name, *Rubripirellula obstinata*, due to its difficulty to grow in culture.

The analysis of huge proteins in the genomes of the three planctomycetes suggests their cell wall or extracellular localization, their potential roles in adhesion and biofilm formation, and their action against stress agents in the complex biofilm of macroalgae.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2017.10.007.

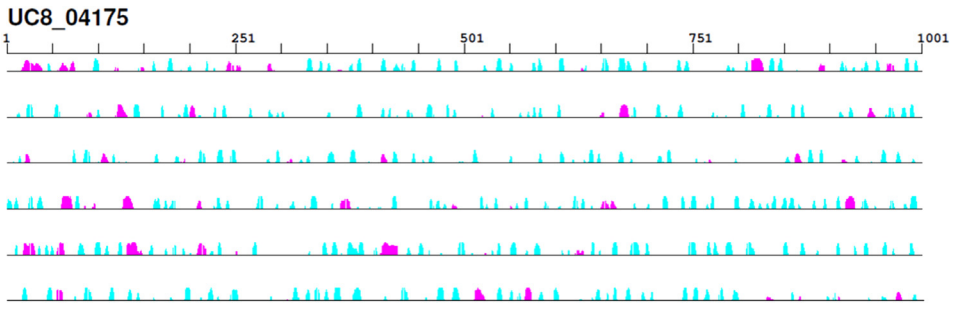


Fig. 3. Predicted secondary structure for UC8_04175. The black horizontal lines represent the sequence of the protein. The predicted α-helices (magenta) and β-strands (cyan) are indicated by bars above each line. The height of the bars is proportional to the confidence of the prediction (Psipred, [78]). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

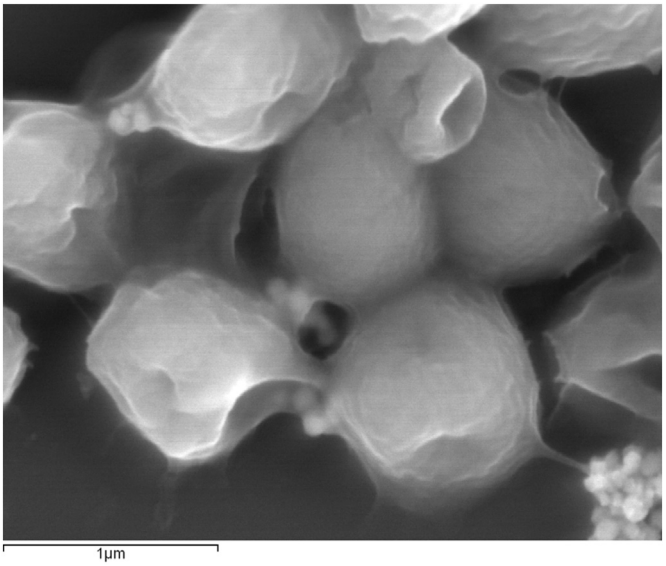


Fig. 4. Scanning electron micrograph of strain FC18^T showing extracellular polymeric substance (EPS) among the cells.

Acknowledgements

This research was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT – Foundation for Science and Technology and European Regional Development Fund (ERDF), in the framework of the program PT2020. NB is funded by Marie Curie ITN FP7-ITN316723-PerFuMe and DPD by the C2A grant EE: 2013/2506 from the Andalusian government. We acknowledge Caitlin Lee Carpenter for her help in proof-reading this article.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the experiments: OML, JH. Performed the experiments: MF,NB, JK. Analyzed the results: MF, NB, DD, OML. Wrote the manuscript: NB, DD, JH, OML. All authors have read and approved the manuscript. We thank the Max Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) for performing the genome sequencing in this study.

Conflict of interest

The authors declare that there is no conflict of interest.

Data deposition

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accessions **LWSI00000000**, **LWSJ00000000** and **LWSK00000000**. The version described in this paper is version **LWSI01000000**, **LWSJ01000000** and **LWSK01000000** respectively.

References

[1] M. Wagner, M. Horn, The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance, Curr. Opin. Biotechnol. 17 (2006) 241–249.
[2] O.M. Lage, J. Bondoso, Planctomycetes and macroalgae, a striking association, Front. Microbiol. 5 (2014) 267.
[3] F.O. Glöckner, M. Kube, M. Bauer, H. Teeling, et al., Complete genome sequence of the marine planctomycete *Pirellula* sp. strain 1, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 8298–8303.
[4] N. Ward, J.T. Staley, J.A. Fuerst, S. Giovannoni, H. Schlesner, E. Stackebrandt, The order *Planctomycetales*, including the genera *Planctomyces*, *Pirellula*, *Gemmata* and *Isosphaera* and the *Candidatus* genera Brocadia, Kuenenia and Scalindua, in: M. Dworkin, S. Falkow, E. Rosenberg, K.H. Schleifer, E. Stackebrandt (Eds.), The Prokaryotes: A Handbook on the Biology of Bacteria, Vol. 7 Springer, NewYork, 2006, pp. 757–793.
[5] M. Pilhofer, K. Rappl, C. Eckl, A.P. Bauer, W. Ludwig, K.H. Schleifer, et al., Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla *Verrucomicrobia*, *Lentisphaerae*, *Chlamydiae*, and *Planctomycetes* and phylogenetic comparison with rRNA genes, J. Bacteriol. 190 (2008) 3192–3202.
[6] O.M. Lage, J. Bondoso, A. Lobo-da-Cunha, Insights into the ultrastructural morphology of novel Planctomycetes, Antonie Van Leeuwenhoek 104 (2013) 467–476.
[7] R. Santarella-Mellwig, S. Pruggnaller, N. Roos, I.W. Mattaj, D.P. Devos, Three-dimensional reconstruction of bacteria with a complex endomembrane system, PLoS Biol. 11 (2013) e1001565.
[8] M.C.F. van Teeseling, S. Neumann, L. van Niftrik, The anammoxosome organelle is crucial for the energy metabolism of anaerobic ammonium oxidizing bacteria, J. Mol. Microbiol. Biotechnol. 23 (2013) 104–117.
[9] T.G. Lonhienne, E. Sagulenko, R.I. Webb, K.C. Lee, J. Franke, D.P. Devos, et al., Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus*, Proc. Natl. Acad. Sci. U. S. A. 107 (2010) 12883–12888.
[10] R. Santarella-Mellwig, J. Franke, A. Jaedicke, M. Gorjanacz, U. Bauer, A. Budd, et al., The compartmentalized bacteria of the *Planctomycetes-Verrucomicrobia-Chlamydiae* superphylum have membrane coat-like proteins, PLoS Biol. 8 (2010) e1000281.
[11] J.A. Fuerst, The Planctomycetes: emerging models for microbial ecology, evolution and cell biology, Microbiology 141 (1995) 1493–1506.
[12] D.P. Devos, Re-interpretation of the evidence for the PVC cell plan supports a Gram-negative origin, Antonie Van Leeuwenhoek 105 (2014) 271–274.
[13] D.R. Speth, M.C. van Teeseling, M.S. Jetten, Genomic analysis indicates the presence of an asymmetric bilayer outer membrane in planctomycetes and verrucomicrobia, Front. Microbiol. 3 (2012) 304.
[14] O. Jeske, et al., Planctomycetes do possess a peptidoglycan cell wall, Nat. Commun. 6 (2015) 7116.
[15] Teeseling MCF, et al., Anammox Planctomycetes have a peptidoglycan cell wall, Nat. Commun. 6 (2015) 6878.
[16] O.M. Lage, J. Bondoso, *Planctomycetes* diversity associated with macroalgae, FEMS Microbiol. Ecol. 78 (2011) 366–375.
[17] J. Bondoso, L. Albuquerque, M.F. Nobre, A. Lobo-da-Cunha, M.S. da Costa, O.M. Lage, *Roseimaritima ulvae* gen. Nov., sp. nov. and *Rubripirellula obstinata* gen. nov., sp. nov. two novel planctomycetes isolated from the epiphytic community of macroalgae, Syst. Appl. Microbiol. 38 (2015) 8–15.
[18] O.M. Lage, L. Albuquerque, A. Lobo-da Cunha, M.S. da Costa, *Mariniblastus fucicola* gen. nov., sp. nov. a novel planctomycete associated with macroalgae, Int. J. Syst. Evol. Microbiol. 67 (2017) 1571–1576.
[19] H. Izumi, E. Sagulenko, R.I. Webb, J.A. Fuerst, Isolation and diversity of planctomycetes from the sponge *Niphates* sp., seawater, and sediment of Moreton Bay, Australia, Antonie Van Leeuwenhoek 104 (2013) 533–546.
[20] J. Bondoso, V. Balagué, J.M. Gasol, O.M. Lage, Community composition of the Planctomycetes associated with different macroalgae, FEMS Microbiol. Ecol. 88 (2014) 445–456.
[21] J. Bondoso, F. Goday-Vitorino, V. Balagué, J.M. Gasol, J. Harde, O.M. Lage, Epiphytic Planctomycetes communities associated with three main lineages of macroalgae, FEMS Microbiol. Ecol. 93 (2017) fiw255.
[22] D.J. Lane, 16S/23S rRNA sequencing, in: E. Stackebrandt, M. Goodfellow (Eds.), Nucleic Acid Techniques in Bacterial Systematics, John Wiley and Sons, New York, 1991, pp. 115–175.
[23] M.P. Cox, D.A. Peterson, P.J. Biggs, SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data, BMC Bioinf. 11 (2010) 485.
[24] M.R. Crusoe, H.F. Alameldin, S. Awad, et al., The khmer software package: enabling efficient nucleotide sequence analysis, F1000Research 4 (2015) 900.
[25] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.
[26] A. Bankevich, S. Nurk, D. Antipov, A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.
[27] Y. Peng, H.C.M. Leung, S.M. Yiu, F.Y.L. Chin, IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, Bioinformatics 28 (2012) 1420–1428.
[28] M. Kearse, et al., Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data, Bioinformatics 28 (2012) 1647–1649.
[29] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics 30 (2014) 2068–2069.
[30] L. Li, C.J. Stoeckert, D.S. Roos, OrthoMCL: identification of ortholog groups for eukaryotic genomes, Genome Res. 13 (2003) 2178–2189.
[31] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, et al., InterProScan 5: genome-scale protein function classification, Bioinformatics 30 (2014) 1236–1240.
[32] A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón, M. Robles, Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676.
[33] Yu NY, et al., PSORTb 3.0: improved protein subcellular localization prediction

with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26 (2010) 1608–1615.
[34] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (3) (2001) 567–580.
[35] M. Galardini, E.G. Biondi, M. Bazzicalupo, A. Mengoni, CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes, Source Code Biol. Med. 6 (2011) 11.
[36] S. Assefa, T.M. Keane, T.D. Otto, C. Newbold, M.A.B.A.C.A.S. Berriman, Algorithm-based automatic contiguation of assembled sequences, Bioinformatics 25 (2009) 1968–1969.
[37] A. Darling, et al., Mauv: multiple alignment of conserved genomic sequence with rearrangements, Genome Res. 14 (2004) 1394–1403.
[38] S. Kurz, et al., Versatile and open software for comparing large genomes, Genome Biol. 5 (2004) R12.
[39] Y. Zhou, Y. Liang, K.H. Lynch, J.J. Dennis, D.S. Wishart, PHAST: a fast phage search tool, Nucleic Acids Res. 39 (2011) W347–W352.
[40] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[41] F. Sievers, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539.
[42] A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, G.J. Barton, Jalview version 2—a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 (2009) 1189–1191.
[43] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Syst. Biol. 59 (2010) 307–321.
[44] J. Bondoso, J. Harder, O.M. Lage, *rpoB* gene as a novel molecular marker to infer phylogeny in Planctomycetales, Antonie Van Leeuwenhoek 104 (2013) 477–488.
[45] M. Guo, et al., Genomic evolution of 11 type strains within family *Planctomycetaceae*, PLoS One 9 (2014) e86752.
[46] D.H. Parks, M. Imelfort, C.T. Skennerton, P. Hugenholtz, G.W. Tyson, Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, Genome Res. 25 (2014) 1043–1055.
[47] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D.A. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science 315 (2007) 1709–1712.
[48] A.Y. Mulkidjanian, K.S. Makarova, M.Y. Galperin, E.V. Koonin, Inventing the dynamo machine: the evolution of the F-type and V-type ATPases, Nat. Rev. Microbiol. 5 (2007) 892–899.
[49] J.S. Lolkema, Y. Chaban, E.J. Boekema, Subunit composition, structure, and distribution of bacterial V-type ATPases, J. Bioenerg. Biomembr. 35 (2003) 323–335.
[50] J.E. Butler, N.D. Young, D.R. Lovley, Evolution of electron transfer out of the cell: comparative genomics of six *Geobacter* genomes, BMC Genomics 17 (11) (2010) 40.
[51] H. Schlesner, *Pirella marina* sp. nov., a budding, peptidoglycan-less bacterium from brackish water, Syst. Appl. Microbiol. 8 (1986) 177–180.
[52] P. Hirsch, M. Müller, *Planctomyces limnophilus* sp. nov., a stalked and budding bacterium from freshwater, Syst. Appl. Microbiol. 6 (1986) 276–280.
[53] I. Kulichevskaya, O. Baulina, P. Bodelier, W. Rijpstra, J. Damsté, S. Dedysh, *Zavarzinella formosa* gen. nov., sp. nov., a novel stalked, *Gemmata*-like planctomycete from a Siberian peat bog, Int. J. Syst. Evol. Microbiol. 59 (2009) 357–364.
[54] W.J. Broadgate, G. Malin, F.C. Küpper, A. Thompson, P.S. Liss, Isoprene and other non-methane hydrocarbons from seaweeds: a source of reactive hydrocarbons to the atmosphere, Mar. Chem. 88 (2004) 61–73.
[55] I. Plettner, M. Steinke, G. Malin, Ethene (ethylene) production in the marine macroalga *Ulva* (*Enteromorpha*) *intestinalis* L. (Chlorophyta, Ulvophyceae): effect of light-stress and co-production with dimethyl sulphide, Plant Cell Environ. 28 (2005) 1136–1145.
[56] Z.A. Popper, G. Michel, C. Hervé, D.S. Domozych, W.G. Willats, M.G. Tuohy, B. Kloareg, D.B. Stengel, Evolution and diversity of plant cell walls: from algae to flowering plants, Annu. Rev. Plant Biol. 62 (2011) 567–590.
[57] C.E. Wegner, et al., Expression of sulfatases in *Rhodopirellula baltica* and the diversity of sulfatases in the genus *Rhodopirellula*, Mar. Genomics 9 (2013) 51–61.
[58] M.R. Amezaga, I.R. Booth, Osmoprotection of *Escherichia coli* by peptone is pediated by the uptake and accumulation of free proline but not of proline-containing peptides, Appl. Environ. Microbiol. 65 (1999) 5272–5278.
[59] J.W. Kim, S.H. Brawley, S. Prochnik, M. Chovatia, J. Grimwood, J. Jenkins, K. LaButti, K. Mavromatis, M. Nolan, M. Zane, J. Schmutz, J.W. Stiller, A.R. Grossman, Genome analysis of *Planctomycetes* inhabiting blades of the red alga *Porphyra umbilicalis*, PLoS One 11 (3) (2016) e0151883.
[60] T. Kohn, A. Heuer, M. Jogler, J. Vollmers, C. Boedeker, B. Bunk, P. Rast, D. Borchert, I. Glöckner, H.M. Freese, H.P. Klenk, J. Overmann, A.K. Kaster, M. Rohde, S. Wiegand, C. Jogler, *Fuerstia marisgermanicae* gen. nov., sp. nov., an unusual member of the phylum *Planctomycetes* from the German Wadden Sea, Front. Microbiol. 7 (2016) 2079.
[61] I.H. Lin, M.T. Hsu, C.H. Chang, Protein domain repetition is enriched in Streptococcal cell-surface proteins, Genomics 100 (2012) 370–379.
[62] Q. Zhang, K.S. Wise, Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent *vaa* genes, Infect. Immun. 64 (7) (1996) 2737–2744.
[63] A.J. Sheets, St. Geme III JW. Adhesive activity of the *Haemophilus* cryptic genospecies Cha autotransporter is modulated by variation in tandem peptide repeats, J. Bacteriol. 193 (2011) 329–339.
[64] C. Gravekamp, D.S. Horensky, J.L. Michel, L.C. Madoff, Variation in repeat number within the alpha C protein of group B streptococci alters antigenicity and protective epitopes, Infect. Immun. 64 (1996) 3576–3583.
[65] M. Hijnen, F.R. Mooi, P.G. van Gageldonk, P. Hoogerhout, A.J. King, G.A. Berbers, Epitope structure of the *Bordetella pertussis* protein P.69 pertactin, a major vaccine component and protective antigen, Infect. Immun. 72 (2004) 3716–3723.
[66] Y.R. Ho, C.M. Li, H.P. Su, J.H. Wu, Y.C. Tseng, Y.J. Lin, J.J. Wu, Variation in the number of tandem repeats and profile of surface protein genes among invasive group B streptococci correlates with patient age, J. Clin. Microbiol. 45 (5) (2007) 1634–1636.
[67] J. Lu, K.D. Marjon, C. Mold, T.W. Du Clos, P.D. Sun, Pentraxins and Fc receptors, Immunol. Rev. 250 (2012) 230–238.
[68] M.W. Vetting, S.S. Hegde, J.E. Fajardo, A. Fiser, S.L. Roderick, H.E. Takiff, J.S. Blanchard, Pentapeptide repeat proteins, Biochemistry 45 (2006) 1–10.
[69] A. Joachimiak, T.A. Springer, R.G. Zhang, J.H. Wang, J.H. Liu, H. Jing, J. Takagi, S. Lindgren, Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins, Structure 10 (2002) 1453–1464.
[70] M. Bycroft, A. Bateman, J. Clarke, S.J. Hamill, R. Sandford, R.L. Thomas, C. Chothia, The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease, EMBO J. 18 (1999) 297–305.
[71] A.Y. Roman, F. Devred, V.M. Lobatchov, A.A. Makarov, V. Peyrot, A.A. Kubatiev, P.O. Tsvetkov, Sequential binding of calcium ions to the B-repeat domain of SdrD from *Staphylococcus aureus*, Can. J. Microbiol. 62 (2015) 123–129.
[72] S. Fujimoto, S. Matsunaga, M. Yonemura, S. Uchiyama, T. Azuma, K. Fukui, Identification of a novel plant MAR DNA binding protein localized on chromosomal surfaces, Plant Mol. Biol. 56 (2004) 225–239.
[73] L. Lin, H. Nakano, S. Uchiyama, et al., Crystallization and preliminary X-ray crystallographic analysis of a conserved domain in plants and prokaryotes from *Pyrococcus horikoshii* OT3. Acta Crystallographica section F: structural biology and crystallization, Communications 61 (2005) 414–416.
[74] S. Koskiniemi, J.G. Lamoureux, K.C. Nikolakakis, C. t'Kint de Roodenbeke, M.D. Kaplan, D.A. Low, C.S. Hayes, Rhs proteins from diverse bacteria mediate intercellular competition, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 7032–7037.
[75] M. Aumailley, The laminin family, Cell Adhes. Migr. 7 (2013) 48–55.
[76] I. Lasa, J.R. Penadés, Bap: a family of surface proteins involved in biofilm formation, Res. Microbiol. 157 (2006) 99–107.
[77] O. Reva, B. Tümmler, Think big—giant genes in bacteria, Environ. Microbiol. 10 (2008) 768–777.
[78] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.

Supplementary information for *Planctomycetes* attached to algal surfaces: insight into their genomes

Table S1 - Number of clustered orthologues proteins shared among the strains

	Number of CDS/Proteins		
	LF1	UC8	FC18
Total CDS annotated	5200	5769	5096
Clustered (%)	3690 (70,96%)	3844 (66.63%)	3164 (62,09%)
Paralogues	516	385	290
Unique (%)	1510 (29,04%)	1925 (33,37%)	1932 (37,91%)

Table S2 - Presence of outer membrane biomarkers and peptidoglycan-related proteins in the proteomes of strains FC18, LF1 and UC8

Description	Protein query	Protein candidate	Expect value
MurA	*Planctopirus limnophila* D5SUH8	LF1_02915	2.00E-55
		UC8_00810	2.00E-31
		FC18_05097	2.00E-30
MurB	*Planctopirus limnophila* D5SME4	UC8_03014	2.00E-83
		FC18_03727	4.00E-75
		LF1_01146	2.00E-62
MurC	*Planctopirus limnophila* D5SME5	UC8_04902	2.00E-18
		LF1_01027	1.00E-17
		FC18_01736	5.00E-15

MurD	*Planctopirus limnophila* D5SR60	LF1_01027	4.00E-109
		UC8_04902	2.00E-90
		FC18_01736	7.00E-38
MurF	*Planctopirus limnophila* D5SR59	FC18_05052	5.00E-82
		LF1_01027	1.00E-28
		UC8_04902	2.00E-22
MurG	*Planctopirus limnophila* D5SND0	FC18_00556	5.00E-29
		LF1_02914	2.00E-23
		UC8_04834	5.00E-18
MurJ	*Planctopirus limnophila* D5SNC0	UC8_04241	5.00E-18
		LF1_02420	3.00E-36
		FC18_03113	1.00E-28
MraY	*Planctopirus limnophila* D5STY1	UC8_05125	1.00E-45
		FC18_01665	3.00E-42
		LF1_04125	9.00E-29
MreB	*Planctopirus limnophila* D5SSP8	UC8_02497	0
		FC18_01328	3.00E-171
		LF1_03296	1.00E-106
FtsK	*Planctopirus limnophila* D5SYV8	LF1_00680	0
		UC8_03091	0

		FC18_03107	0
DapA	*Planctopirus limnophila* D5SQS5	LF1_04593	4.00E-25
		UC8_00877	8.00E-22
		FC18_00625	1.00E-13
DapB	*Planctopirus limnophila* D5STR2	FC18_01544	1.00E-96
		UC8_05144	1.00E-93
		LF1_00265	5.00E-88
DapF	*Planctopirus limnophila* D5SNB0	FC18_04869	5.00E-110
		LF1_00400	1.00E-95
		UC8_00843	1.00E-09
DapL	*Planctopirus limnophila* D5SN60	LF1_04723	5.00E-60
		UC8_01465	6.00E-53
		FC18_02158	9.00E-41
TolQ	*Planctopirus limnophila* D5SWI1	LF1_01768	2.00E-42
		UC8_05161	2.00E-32
		FC18_02298	4.00E-27

Description	Protein candidate	Protein query	Expect value
Lipossacharide insertion	UC8_03269	*Rhodopirellula baltica SH 1* NP_867548	0
	LF1_02324		0
	FC18_02068		0
OMP insertion and presence	UC8_04107	*Rhodopirellula baltica SH 1* NP_869683	4.00E-151
	UC8_04106	Gimesia maris DSM 8797 ZP_01854098	8.00E-105
	UC8_04964	*Isosphaera pallida* WP_013564302	8.00E-06
	LF1_03825	*Planctopirus limnophila* WP_013110622	9.00E-101
	LF1_03826	*Rhodopirellula baltica* WH47 EGF26385	0
	FC18_01993	*Gimesia maris* DSM 8797 ZP_01854098	3.00E-112
	FC18_01992	*Blastopirellula marina* DSM 3645 ZP_01088553	1.00E-105
TonB system	UC8_04183	*Blastopirellula marina* DSM 3645 EAQ78342	6.00E-14

Table S3. CRISPR associated proteins in FC18 and LF1

Protein candidate	Putative protein
FC18_03468	CRISPR-associated endonuclease Cas9
FC18_04411	CRISPR-associated protein Cas5
FC18_04412	CRISPR-associated protein (Cas_Csd1)
FC18_04413	CRISPR-associated protein Cas7/Csd2
FC18_04415	CRISPR-associated protein Cas4/endonuclease Cas1 fusion
FC18_04416	CRISPR-associated protein Cas4/endonuclease Cas1 fusion
FC18_04417	CRISPR-associated endoribonuclease Cas2
LF1_00307	CRISPR-associated endonuclease Cas6/Csy4
LF1_00308	CRISPR-associated protein Csy3
LF1_00309	CRISPR-associated protein Csy2
LF1_00310	CRISPR-associated protein Csy1
LF1_00311	CRISPR-associated nuclease/helicase Cas3 subtype I-F/YPEST
LF1_00312	CRISPR-associated endonuclease Cas1
LF1_02821	CRISPR-associated endonuclease Cas9

Table S4 – Vacuolar-type ATPase present in strains UC8^T and LF1^T and closet hits.

Protein	Predicted function	Hhpred closest Hit	Hhpred E-value	Microorganism
UC8_03687	V-type sodium ATPase catalytic subunit A	5bn3_A V-type ATP synthase alpha chain	6,00E-162	*Nanoarchaeum equitans*
UC8_03688	hypothetical protein	1v9m_A V-type ATP synthase sub	7.4E-10	*Thermus thermophilus*
UC8_03689	V-type ATP synthase subunit E	4efa_E V-type proton ATPase	3.6E-27	*Saccharomyces cerevisiae*
UC8_03690	V-type ATP synthase subunit K	3j9t_Y V-type proton ATPase subunit C	9.2E-36	*Saccharomyces cerevisiae*
UC8_03691	V-type ATP synthase subunit I	3j9t_b V-type proton ATPase subunit A	8.4E-86	*Saccharomyces cerevisiae*
UC8_03692	V-type ATP synthase subunit D	3aon_A V-type sodium ATPase subunit D	9.5E-60	*Enterococcus hirae*
UC8_03693	V-type sodium ATP synthase subunit B	2c61_A A-type ATP synthase non-catalytic subunit B	1.7E-110	*Methanosarcina mazei* GO1
LF1_01687	V-type sodium ATPase catalytic subunit A	1v9m_A V-type ATP synthase subunit C	0.00011	*Thermus thermophilus*
LF1_01688	hypothetical protein	2c61_A A-type ATP synthase non-catalytic subunit B	1.7E-110	*Methanosarcina mazei* GO1
LF1_01689	V-type ATP synthase subunit E	3v6i_A V-type ATP synthase subunit E	2.5E-26	*Thermus thermophilus*
LF1_01690	V-type sodium ATP synthase subunit K	3j9t_Y V-type proton ATPase subunit C	1.8E-34	*Saccharomyces cerevisiae*
LF1_01691	V-type ATP synthase subunit I	3j9t_b V-type proton ATPase subunit A	6.8E-90	*Saccharomyces cerevisiae*
LF1_01692	V-type ATP synthase subunit D	3aon_A V-type sodium ATPase subunit D	3.6E-61	*Enterococcus hirae*
LF1_01693	V-type sodium ATPase subunit B	2c61_A A-type ATP synthase non-catalytic subunit B	7.3E-111	*Methanosarcina mazei* GO1

Table S5.

Comparative analyses of sulfatases, polysaccharide degrading enzymes, stress response proteins and cytochromes detected in the proteomes of strains FC18^T, LF1^T and UC8^T

Proteins		FC18	LF1	UC8
Sulfatases		61	36	97
Polysaccharide degrading enzymes	cellulase	1	2	1
	agarase	0	1	1
	amylase	1	1	1
	N-acetylglucosamine deacetylase	1	1	1
	xylanase	9	5	0
	porphyranase	3	3	1
Stress response	Copper related proteins	4	4	8
	Co-Zn-Cd resistance proteins	8	8	12
	Cd transporting ATPase	2	1	3
	Mercury resistance proteins	1	1	1
	Arsenate related proteins	2	1	1
	Multidrug resistance proteins	15	13	22
	SOS response	1	1	1
	Glutathione	3	5	6

	Superoxide dismutase	2	2	3
	Catalase-peroxidase	2	1	2
	UV resistance proteins	10	10	9
	Heat-shock proteins	3	2	4
	Betaine related proteins	2	1	2
	Glycine related proteins	6	12	10
	Proline related proteins	3	3	6
	Trehalose related proteins	11	4	14
	D-tyrosyl-tRNA deacetylase	1	1	1
Cytochrome		47	27	103

Protein candidate	Length (aa)	Localization	Annotation	Identified domain in Pfam
LF1_00233	8275	Extracellular	LF1_00233 Bifunctional hemolysin/adenylate cyclase precursor	PF00353.17 HemolysinCabind domain (28x)
				PF13229.4 Beta_helix domain (5x)
				PF10342.7 GPI-anchored domain
				PF13385.4 Laminin_G_3
LF1_01856	6270	Extracellular	LF1_01856 Bifunctional hemolysin/adenylate cyclase precursor	PF00353.17 HemolysinCabind domain (30x)
				PF12799.5 LRR_4 domain (1x)
				PF13385.4 Laminin_G_3 domain (1x)
				PF17210.1 SdrD_B domain (1x)
UC8_01824	9958	Extracellular	UC8_01824 Bifunctional hemolysin/adenylate cyclase precursor	PF00805.20 Pentapeptide domain (6x)
				PF00353.17 HemolysinCabind domain (24x)
				PF00354.15 Pentaxin domain (3x)
				PF10282.7 Lactonase domain (2x)
				PF00801.18 PKD domain (2x)
UC8_03858	7047	Extracellular	UC8_03858 Bifunctional hemolysin/adenylate cyclase precursor	PF00353.17 HemolysinCabind domain (32x)
				PF00801.18 PKD domain (1x)
LF1_03591	6494	Extracellular	LF1_03591 Matrixin	PF00413.22 Peptidase_M10 domain (1x)
				PF04151.13 PPC domain (1x)
UC8_04175	6003	Extracellular	UC8_04175 Matrixin	PF00413.22 Peptidase_M10 domain (1x)
UC8_01761	6601	Extracellular	UC8_01761 tRNA(Glu)-specific nuclease WapA precursor	PF04151.13 PPC domain (1x)
				PF13385.4 Laminin_G_3 domain (1x)
				PF13205.4 Big_5 domain (1x)
				PF07705.9 CARDB domain (9x)
				PF05593.12 RHS_repeat domain (6x)
LF1_01536	6783	Extracellular	LF1_01536 Exoglucanase B precursor	PF17164.2 DUF5122 domain (10x)
				PF00801.18 PKD domain (1x)
				PF00041.19 fn3 domain 83x)
FC18_01343	7202	Periplasmic	FC18_01343 Calx-beta domain protein	PF14252.4 DUF4347 domain (1)
				PF16184.3 Cadherin_3 domain (1x)
				PF03160.12 Calx-beta domain (19x)
				PF17210.1 SdrD_B domain (1)

FC18_02694	7782	Periplasmic	FC18_02694 Serine-aspartate repeat-containing protein D precursor	PF14252.4 DUF4347 domain (1x) PF17210.1 SdrD_B domains (27x)
FC18_05157	7270	Periplasmic	FC18_05157 Laminin G domain protein	PF14252.4 DUF4347 domain (1) PF17164.2 DUF5122 domain (13x) PF13385.4 Laminin_G_3 domain (3x) PF16184.3 Cadherin_3 domain (3x)
FC18_03125	7644	Periplasmic	FC18_03125 hypothetical protein	PF14252.4 DUF4347 domain (1x) PF01345.16 DUF11 domain (1x)
UC8_04174	6091	Extracellular	UC8_04174 hypothetical protein	PF04151.13 PPC domain (3x)
LF1_01999	6214	Periplasmic	LF1_01999 hypothetical protein	No domain match

Table S8. Domains and subcellular localization predictions for the huge proteins.

4.3

ICBdocker: a Docker image for proteome annotation and visualization

Bioinformatics, 2018, 1–2
doi: 10.1093/bioinformatics/bty493
Advance Access Publication Date: 19 June 2018
Applications Note

Sequence analysis

ICBdocker: a Docker image for proteome annotation and visualization

Nicola Bordin* and Damien P. Devos*

Laboratory of Evolutionary Innovations, Centro Andaluz de Biologia del Desarollo, CSIC, Universidad Pablo de Olavide, Carretera de Utrera, Seville 41013, Spain

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 26, 2018; revised on May 22, 2018; editorial decision on June 13, 2018; accepted on June 15, 2018

Abstract

Summary: We introduce ICBdocker, a Docker environment that allows the annotation of functional and structural features of proteomes through a Python/Perl pipeline. DataTables pages make it easy to set up a web-resource for research groups with a focus on the same organisms or datasets. The results are available as tab-separated values files and HTML, allowing data analysis and browsing. The pipeline focuses on modularity and scalability, with capability of integrating with multiprocessing and high-performance computing clusters.

Availability and implementation: ICBdocker is freely available on DockerHub at https://hub.docker.com/r/bordin89/icb/ Source code and documentation are available on GitHub at: https://github.com/bordin89/ICB_docker.

Contact: bordin89@gmail.com or damienpdevos@gmail.com

1 Introduction

Thanks to advancements in DNA sequencing, various consortiums and laboratories can now obtain an organism's genome. However, they often lack the means for protein functional annotation. Likewise, advances in '-omics' methods allow for the determination of sets of proteins, often with limited derived functional information. ICBdocker addresses three main issues related to this problem. First, the majority of current web resources in biology are focused on model organisms, leaving laboratories working with newly sequenced or non-model organisms with a lack of functional annotation. Second, many annotation pipelines are either too focused on a specific protein aspect (interactions, location, expression) or their usage requires some computational expertise. This is because assigning function to a protein sequence is a considerable task that requires the integration of multiple sources on function such as Gene Ontologies, domains, secondary and tertiary structure features. To address such issues, we have recently introduced an Integrative Computational Biology (ICB) pipeline for the simultaneous consideration of multiple functional aspects (Bordin *et al.*, 2018). An application of ICB is available at the PVCdb section of PVCbase (http://pvcbacteria.org/pvcbase/). Finally, containerization is a novel way to distribute software that runs inside a 'container' in an operating system. It allows the user to run a Linux environment in a Microsoft Windows or MacOS computer, with a lack of interference with previously installed software on the host computer. In computational biology, it can be used to deploy data analysis environments with multiple predictors that are otherwise time-consuming to set up. Containerization tools such as Docker are of pivotal interest in research reproducibility, since it allows experiments to be replicated using the same environment (software, datasets and conditions; Boettiger, 2015).

2 Method overview

ICBdocker is a Python/Perl pipeline that, provided with one or more protein sequences, performs multiple analyses focused on different protein functional descriptions. The pipeline can run a variable set of modules at once based on user needs. Homology search is performed using PSIBLAST (Altschul *et al.*, 1997) against UniProtKB/SwissProt (Bairoch *et al.*, 2004). The hits matching an e-value of 1e–3 and a query coverage of more than 75% is then parsed for Gene Ontology (GO) entries, keywords and Enzyme Commission (EC) numbers (Ashburner *et al.*, 2000). InterProScan (v5.16-55.0; Jones *et al.*, 2014; Hunter *et al.*, 2009) searches for

protein signatures on several databases, including PFAM, PANTHER and SUPERFAMILY among others and results are parsed for KEGG-pathway entries (Ogata *et al.*, 1999) and additional GO terms. Tertiary structure prediction based on homology is performed using HHsuite (Remmert *et al.*, 2012). HHblits is used to create a multiple sequence alignment (MSA) by comparing the query Hidden Markov Model (HMM) to the UniProt20 HMMs database with the option–addss which adds secondary structure information using PSIPRED (McGuffin *et al.*, 2000). The resulting MSA is converted in a HMM using hhmake and searched in the pdb70 database (release 14Sept16; Berman *et al.*, 2002). The output of the search is parsed with a minimum threshold of 1e–3 and mapped on SIFTS (Velankar *et al.*, 2012) to obtain GOs and EC numbers. Raw alignments are kept for eventual modelization using MODELLER (Sali and Blundell, 1993). The presence of signal peptides, transmembrane helices and disorder are determined through modules and parsers of SignalP4.1, TMHMM and IUPred (Dosztányi *et al.*, 2005; Krogh *et al.*, 2001; Petersen *et al.*, 2011). The modules results are collected, ordered and summarized to generate a tabular-separated values file and a HTML page preformatted using a jQuery DataTables plugin. The plugin adds sorting, paging and filtering to plain HTML tables. All results, including the raw data from the predictors, are kept for further analyses.

3 Implementation

ICBdocker is provided through a Docker image that can be pulled from DockerHub (Merkel, 2014). The image runs on every architecture supported by the Docker engine. This allows seamless download and installation of all the pipeline dependencies and databases. A shared data folder can be passed to the container alongside the parameters for multi-core processing. The image can run in a queue system like SGE or SLURM with multiple instances and doesn't conflict with previously installed software or databases. ICBdocker was used to characterize 39 proteomes of relevant PVC (Planctomyces-Verrucomicrobia-Chlamydia) bacterial strains and the results obtained, including the raw data, are available at PVCdb (Bordin *et al.*, 2018).

4 Conclusions

ICBdocker provides easy deployment of a computational pipeline for protein annotation, including its required databases. The analyses performed are visualized through DataTables offering a global overview of protein features. The output formats were designed to be easily implemented in web resources for shared analysis.

Acknowledgements

The authors thank Juan Carlos González-Sánchez for creating the earliest core of ICB. They also thank Caitlin Lee Carpenter for her help in proofreading this manuscript.

Funding

This work was supported by the Marie Curie Innovative Training Network [FP7-ITN316723- PerFuMe] to NB, by the Andalusian Government C2A grant [EE: 2013/2506] and by the Spanish Ministry of Economy and Competitiveness (BFU2013-40866-P, BFU2016-78326-P) to DPD.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.*, **5**, 39–55.

Berman,H.M. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.

Boettiger,C. (2015) An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.*, **49**, 71–79.

Bordin,N. *et al.* (2018) PVCbase: an integrated web resource for the PVC bacterial proteomes. *Database*, bay042.

Dosztányi,Z. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Hunter,S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

Jones,P. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

McGuffin,L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.

Merkel,D. (2014) Docker: lightweight Linux containers for consistent development and deployment. *Linux J*, http://doi.org/10.1097/01.NND.0000320699.47006.a3.

Ogata,H. *et al.* (1999) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Bio.*, **234**, 779–815.

Velankar,S. *et al.* (2012) SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.

Supplementary material for ICBdocker: a Docker image for proteome annotation and visualization

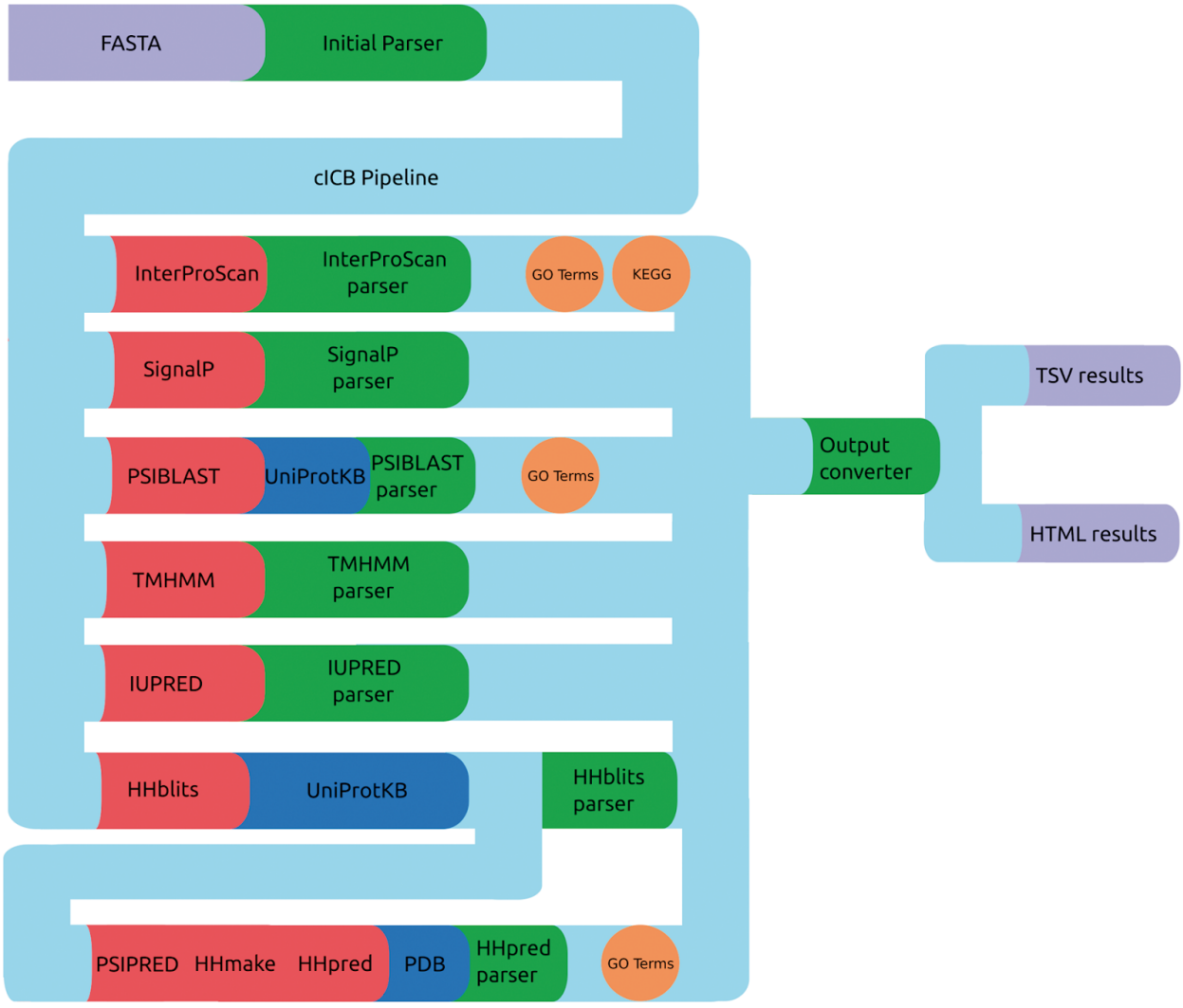


Figure 14. ICB pipeline structure. Files (purple), tools (red), databases (blue) and scripts (green).

Extensive ICB modules description:

PSIBLAST

The PSIBLAST module performs a homology search on UniProt/SwissProt, which provides information on similar hits with their Gene Ontology (GO), keywords, and Enzyme Commission (EC) numbers (Altschul et al., 1997).
Gene Ontology provides information on the role of the protein in the cell. In particular on the molecular function it performs, the localization in the cell where it carries out its processes, and the overall biological processes in which it is involved. Enzyme Commission numbers provide the specific chemical process the protein performs.

HHBlits on UniProt

HHblits (Remmert et al., 2011) maps sequence domains on the protein by creating a multiple sequence alignment based on the UniProt20 Hidden Markov Model Database. Secondary structure prediction features (like alpha-helices and beta-sheets) are added using PSIPRED (McGuffin et al., 2000), increasing the sensitivity of HHblits.

HHpred on PDB

HHpred consists of several tools that allow it to map similar structures from PDB on the protein under examination (Söding, 2005). This module provides information on function, GOs, and EC numbers. As a secondary advantage, the raw alignments generated by HHpred allow the user to create a homology-based model of the protein's structure using other tools like MODELLER (Sali and Blundell 1993).

InterProScan

InterProScan searches for protein signatures on InterPro's several databases, including PFAM, PANTHER, and SUPERFAMILY, among others (Jones et al., 2014, McDowall and Hunter, 2011). This module elucidates functional domains on the protein and provides information on the protein's role in the cell through KEGG-pathway entries and additional GO terms (Ogata et al., 1999).

SignalP

SignalP 4.1 predicts the presence or absence of a signal peptide at the N-terminus of the sequence, indicating if the protein is excreted from the cell or expressed in the cytosol (Petersen et al, 2011).

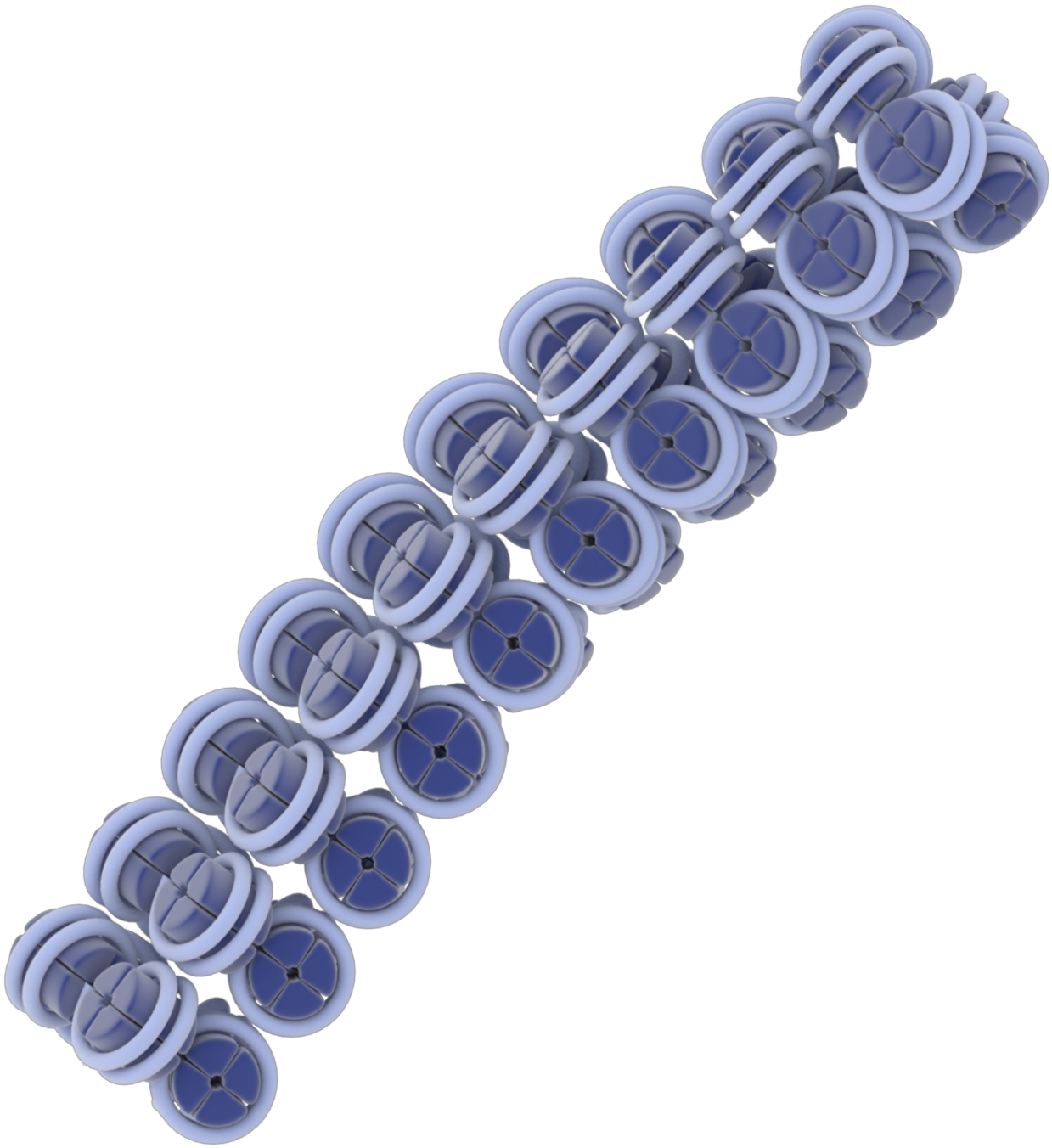
IUPRED

IUPRED predicts the average level of disorder from the protein sequence. The tool outputs a score for each amino acid. A score above 0.5 indicates a disordered residue, while a score below that threshold indicates a residue in a folded region. Using these values, the module gives information on contiguous globular portions of the protein. (Dosztányi et al., 2005).

TMHMM

Using TMHMM, the module is able to predict the presence and the localization of TransMembrane Helices (TMHs) in the protein sequence, indicating which portions are extracellular or cytoplasmic (Krogh et al, 2001).

5 Discussion



5.1 Information integration is the key to a better protein function prediction

Proteins of unknown function (PUFs) and the need for a better, more complete analyses of individual proteins are a recurrent problem in all large-scale analysis, including genomics and system biology. Many strategies and tools, either computational or experimental, have been designed to assign a function to proteins. More than three decades of computational analysis have addressed this issue and this accumulated knowledge forms the basis of protein bioinformatics. Computational tools have been developed to go beyond sequence-only information (i.e. BLAST) and assign function to a protein based on alternative information, such as structure, genome context, domains, interaction, etc.. One of the problems with most of those computational methods is that they have been applied to artificial test sets, leaving the bias towards related proteins and how they relate to 'real' genome unclear. Another problem is that most have been applied in isolation without consideration of alternative information. In contrast, many large-scale experimental assays have interrogated complete genomes but with limited reference to advanced computational information, most of the time annotating proteins based on limited homology inferences. Where these two worlds meet and how they complement each other, especially when applied jointly on a particular genome, is unclear.

Integration of information and resources about protein function to obtain a consensus in protein function prediction is therefore paramount. An example of data integration in protein function prediction is the manual annotation process that is performed for the UniProt/SwissProt database. Every deposited protein undergoes quality checks on its sequence, removing discrepancies and redundant entries due to alternative splicing and frameshift. The integrative approach is applied by using several predictors that determine domains, TMHs, subcellular localizations, GOs, and other protein features. All this information is considered to obtain a reliable annotation. The SwissProt curators complement and assess this by merging data from other sources besides *in-silico* predictors, such as literature curation, expression evidence, and structure-based methods.

While being extremely effective for a limited number of proteins, this approach is no longer sustainable for the amount of proteins sequences that are deposited every year. One of the drawbacks of SwissProt is that these annotations can't be browsed for multiple proteins at once, making it difficult to find a common theme in a set of proteins. The Integrative Cell Biology approach we have developed allows researchers to observe a complete set of annotations for multiple proteins at the same time.
The ultimate goal for ICB is to use this integrative approach, automate it, and allow multiple queries simultaneously to have an overview of a larger dataset.

5.2 The future of protein function prediction

Since 2005 at the Intelligent Systems for Molecular Biology (ISMB), the major conference of computational biologists worldwide, the protein function prediction community has gathered in the Automated Function Prediction (AFP) Special Interest Group (AFP/SIG) to discuss and assess the current status of function prediction tools and present novel ways to address issues. Out of the AFP/SIG, the Critical Assessment of Function Annotation (CAFA) was born in 2010 as a community experiment to evaluate the quality of annotation tools on a predefined dataset of thousands of uncharacterized proteins (Radivojac et al, 2013). Since the first edition, CAFA has been one of the driving forces behind the improvement of protein function prediction and attending the meeting is a great way to keep up with the latest trends in the field.
Recent approaches to CAFA, then released for general protein annotation, use a combination of different novel techniques, in particular, Deep Learning, evolution and structure-based methods.
Since the first CAFA experiment, most predictors implement Machine Learning approaches, which consist of creating a curated dataset of proved "sequence to function" relationships and using it to train a classifier algorithm that automates annotation. One of the issues of these early methods is that human intervention was still required, and the machine learning process was just assigning weights to a manual selection of features. Machine learning algorithms have been used for detecting sequence features since the

early 2000s (Keşmir et al., 2002). More recently a novel subset of machine learning algorithms, called Deep Learning, have been implemented in protein function predictors. Deep Learning algorithms mimic the thought process of the human brain, using layered algorithms in order to create an artificial neural network (ANN) that can learn and make decisions on its own. Applied to proteins, ANN are able to extract information from raw sequences without human intervention. At every iteration of the algorithm, the classification improves, showing promising results in terms of speed and precision.

Evolution-based methods have been used recently to assess the quality of annotations. Techniques such as genomic phylostratigraphy and protein architecture prediction overcome some of the limitations of homology-based methods and allow us to trace the evolution of a protein and its presence back to a common ancestor (Domazet-Lošo T et al., 2007).

The integration of these methods in automated function predictors increased their accuracy, and from CAFA1 in 2010 to CAFA2 in 2013-2014, the top methods showed encouraging progress. However, raw scores from the competition indicate there is still room for improvement (Jiang et al., 2016). Current CAFA evaluations are based on the correct prediction of Gene Ontologies, which has limitations. Tools being specifically built for the task are mostly based on sequence and structure methods, biasing the results towards an *in-silico* only approach and omitting any wet-lab based results.

5.3 Towards the integration of in-vivo and in-silico

While Bioinformatics is devoted to the development of tools aiding biological analyses, Computational Biology is a much more intriguing field. Applying computational techniques to biological datasets allows us to make biological discoveries that wouldn’t be possible using only wet lab experiments.

The integration of in-vivo and in-silico information is the key to answering ongoing questions in cell biology, evolution, and medicine.

Most of the techniques developed to resolve pathways and identify orphan enzymes rely on the integration of a few different sources for pathway complementation. Genomic contexts across multiple species (Smith et al., 2012, Green and Karp 2007), genes

positions and phylogenetic profiling (Yamanishi et al., 2007), and graph analyses (Ye et al., 2005) are all viable approaches to identify missing enzymes but are both limited and limiting in integrative terms. Other methods integrate some of the previous approaches in order to obtain a consensus but lack wet-lab data (Kharchenko et al., 2006, Zhu et al., 2012).

The Sali group recently used a combination of chemoinformatics, genomic context, virtual screening, and ligand-binding analysis to predict the L-gulonate catabolic pathway in *Haemophilus influenzae* Rd KW20. The strengths of this method are the generalizing of information associated to a pathway instead of relying on a single source of information and predicting gaps in annotations based on the lack of biochemical knowledge instead of its presence.

This particular approach, backed by subsequent results confirmation by enzymology, crystallography, and metabolomics proved that systems biology and structural biology are not separate fields, but complementary, and all sources of information should be used to close knowledge gaps by using an integrative approach (Calhoun et al., 2018) (**Figure 15**).

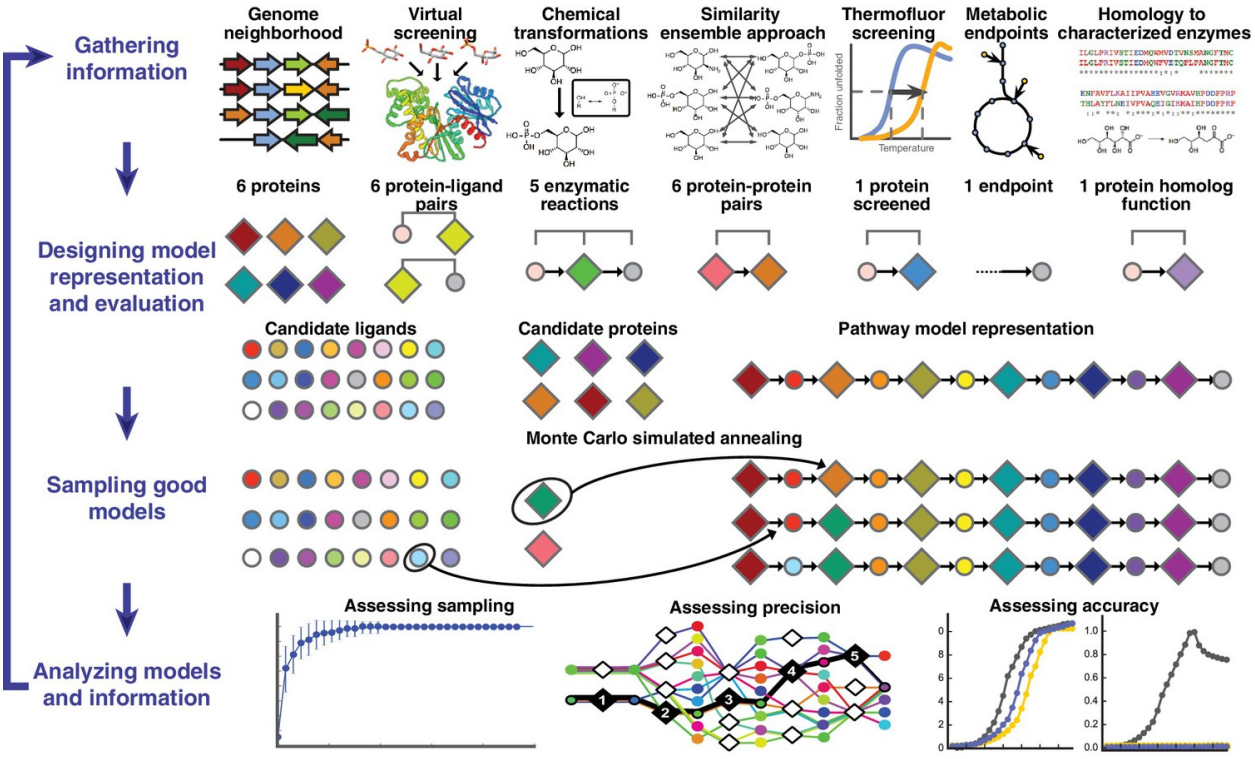


Figure 15. Overview of integrative pathway mapping method. (Calhoun et al., 2018)

Cell imaging can also play a part in elucidating cell features, such as the exocyst structure. Live cell imaging in yeast and fluorescent chromophores were used to measure the distances between the extremities of proteins of a macromolecular complex. The distances were used as trilateralization constraints and modeled in the 3D space, generating a model that was coherent with the previous knowledge of the structure and features of the exocyst complex (Picco et al., 2017).
These are just a few examples of how integrating in vivo and in silico analyses has proven to be a successful approach to solving biological questions.

5.4 Solving the problem of protein "darkness"

Despite all the past and current efforts, the percentage of "putative", "uncharacterized", or proteins of unknown function is still staggeringly high. Some of these proteins can be annotated using the aforementioned integrative approaches, but this is not always possible. Some limitations in computational biology are due to the current insufficient knowledge we have of proteins. We can begin to understand a protein's function if we know the role the protein has in the cell, its molecular activity, and/or where it's expressed. Obtaining knowledge of a protein family that is currently unknown requires biological experiments, such as expression, purification, and localization studies. Once the existence of the protein is confirmed, there are potential problems with further characterization. While some of these can be characterized at the sequence level, some others have diverged enough that prediction tools aren't decisive. Therefore, the next step is to characterize the protein's three-dimensional structure. However, sometimes obtaining the protein structure through crystallography or Nuclear Magnetic Resonance (NMR) is not feasible. Proteins that can't be characterized structurally usually fall into two categories: "known unknowns" and "unknown unknowns". "Known unknowns" are proteins with a high level of intrinsic disorder or with transmembrane helices encoded in its sequence, making them hard to crystallize. "Unknown unknowns" are proteins that still have to be characterized but can be crystallized and aren't describable by other means. Both categories contribute to what is known as the "dark matter" of the protein universe, something we know is there, but can't shed light upon it (Levitt, 2009). The dark proteome

potentially holds a treasure trove of proteins that have new functions and folds (Perdigao et al., 2015). Dark proteins are ubiquitous in the tree of life, they aren't particularly disordered and are present in higher percentages in Metazoa and viruses (Bordin et al., manuscript in preparation). As CAFA was born to improve protein annotation methods, in 1994 the Critical Assessment of protein Structure Prediction (CASP) (Moult et al., 1995) was created to assess protein structure prediction methods. Ideally, the two consortiums would join their efforts to unveil the features of dark proteins.

5.5 ICBdocker advantages, disadvantages and future prospects

The ICB pipeline and its Docker container, ICBdocker, has its advantages and disadvantages, like any other piece of scientific software. The container allows the user to easily deploy a fully functional computational pipeline for annotating whole proteomes. The programs included cover most of the aspects that are useful for assigning a protein's function, and its modularity guarantees scalability and expansion to include novel tools adept for implementation in high performance computing facilities. The DataTables plugin included in the HTML output generates a smart table that allows filtering and extracting subsets of proteins based on keywords, as well as paging and sorting. Being a fully functional web page, it can be implemented in web servers and create shared resources for research communities. This containerized approach is new in protein bioinformatics, since providing web resources through containers has only been used in the genomics community so far, through tools like GenomeHubs (Challis et al., 2017). Although PVCbase and ICBdocker have proved useful for the community (over 150 single users from 8 different countries in less than a year) and focuses mostly on user-friendliness, it has plenty to improve with time.

When developing ICBdocker, I faced a choice. From a resource management standpoint, ICBdocker would have been better designed if the tools were separated from the databases, making it more flexible and smaller in size. A user could download just the container with the modules they were interested in using, while in the current configuration the container is monolithic. The pipeline code itself is modular, so separating the modules

in different containers wouldn't be difficult, but this flexibility comes with a cost to the user. The pipeline was designed for wet-lab biologists with little to no experience in setting up these environments and splitting the pipeline into its core components defeats the purpose of its creation.

Containers, by definition, are closed and almost airtight. While they were designed to be this way to avoid disruption in a production environment, it means containers are isolated and cannot communicate between themselves or outside sources. By default, Docker containers don't communicate with the Internet, and creating a networking interface for them requires some alterations at the OS level by mapping or opening ports, which is beyond most users' capabilities. In bioinformatics, plenty of tools are available as web services through REST APIs, allowing users to run analyses remotely on a hosting server and retrieve the results locally. Some tools have databases so massive that the required computational power renders their use impossible on a regular laptop, and APIs are the only way to obtain results from these tools. Containers' access to data is limited by their networking issues and do not include results from remote services, making broader analyses difficult.

If the issues with container setups and networking are eventually fixed, the best scenario would be a pipeline made entirely of containers, with the possibility to run them separately and the ability to use different versions of databases as needed.

One major downside of containerization is Docker itself. Docker is a proprietary technology, with closed and proprietary source code. The central repository for Docker images, DockerHub, hosts plenty of scientific software that was funded by taxpayers and shouldn't be residing with a private hosting company. If Docker one day decides to remove these tools from its repository, it would be a blow to the scientific community. Likewise, if the company someday fails, a large body of knowledge could disappear with it.

When Docker first introduced its technology, it acknowledged the risk of having only a centralized repository for all available containers, so it released portions of its source code to the public. Containers are now the backbone of every major tech company, such as Google, Apple, and Amazon. Most players in the technology field joined forces in the Open Container Initiative (OCI) to create an open source container solution for both

regular users and enterprise. While the first products, such as runC and the OS for managing containers like CoreOS and Kubernetes, are mostly enterprise-oriented and are used on cloud hosting or HPC platforms, other solutions for smaller initiatives are being developed. In Science, the ELIXIR consortium was founded in 2014 to merge and regulate multiple efforts to build a distributed infrastructure for life-science information. Among its members, there are teams devoted to creating alternatives to current containerization programs and define the best practices for their use in reproducible research. ELIXIR partners already host a great variety of services and databases related to DNA and proteins. Hopefully, we will have soon an open, publicly-funded alternative to Docker and DockerHub for container creation and hosting.

5.6 Integrative Cell Biology: what's next

The concept behind Integrative Cell Biology isn't limited to its current form of structure and sequence methods combined in a computational pipeline. In its next iteration, it can be improved in several areas to follow the latest trends in the field. Aspects where the program could be improved are in its methodology, visualization, scope, and development.

Pipeline architecture and additions

The main program of ICB consists of a series of Python scripts that allows the user to launch the different predictors internally, without the need to call the single programs or parsers directly. This is managed by command-line flags passed to the main script. The architecture of the pipeline is modular, so the inclusion of new sources of information can be done with few modifications to the code by adding the corresponding flag to the ICB application. Some improvements on the algorithm design would be ideal. The first "beta" version of the pipeline consisted of monolithic programs that treated each module as subroutines, and any further module addition required amending the main program. A good approach for the next version of ICB would be to break the main program into single components and treat the embedded modules as external libraries or functions not

residing in the main script. This would result in a more streamlined code, with a smaller chance of making the pipeline unfunctional if the wrong line of code gets modified by mistake. With projects involving large amounts of proteins, this would allow for the modification of the modules without halting the annotation process.

Docker

The previous chapter about the containerization situation pointed out the problematic choice between flexibility and ease of use for the user. A potential solution is to create two versions of the Docker image. One the complete ICB container, with its databases, dependencies, and tools already deployable for less experienced users as a one-click install on large HPCs. The second as a series of containers with each tool and database available as separate resources for more experienced users or when storage space is a concern.

Data visualization

The next version of ICB would benefit from improvements on the graphic representations of the results. As the PVCBlast platform was modified to allow a link-out system to the DataTables' prefiltering, retrieving the annotation for a specific hit, a link-out system for each of the results in the DataTable is necessary. At the moment the GO annotations and the entries from UniProt or PDB are plain text and require the user to open a tab and search for an expanded description on the corresponding website. The next version of the DataTables should provide hyperlinks that automate this task.
If new modules will implement data from genomics and transcriptomics, additional outputs with coordinates and a Genbank annotation file could be of help in further pathway or gene expression analyses.

Future modules

The current version of the ICB pipeline includes only structure and sequence-based methods. Although these are informative about the protein itself, hints of the protein's role in the cell also come from interactions with its environment, when and where it's expressed alongside other proteins in a specific process (Pearson, 2015).
Further modules of ICB might implement information from several new predictors, databases, and alternative sources of information, including upcoming tools that aren't available yet.
In order to create the next version of ICB, we decided to start from a clean slate and plan the architecture of the pipeline accordingly. Its modules will be organized in the following families, each encompassing a source of information and with improved interconnectivity between them.

- Sequence: Most of the current methods in the first version of ICB will be collected in the Sequence module, including homology-based methods, motifs, domain detectors, transmembrane helices, and disorder predictors.
- Structure: This family of modules will include structural domains predictors (HHpred), 3D structural aligners such as MOMA (Guitiérrez et al., 2016), and other secondary structure prediction tools.
- Interactions: Protein interaction databases, such as STRING (Szklarczyk et al., 2015) provide information on protein interactions, allowing researchers to determine which processes the protein is involved in and which cell compartment it is expressed. In the case of STRING, this information is determined through several sources, such as text-mining, proteomics, coexpression, and other experiments. Other approaches, like InterPreTS, use structural information to determine a set of potential interactors (Aloy and Russell, 2003).
- Genome: The integration of gene fusion/fission experiments, as well as other in-silico based features such as genome proximity help to elucidate the function of a protein by looking at neighboring genes. This is particularly true in the case of bacteria, where proteins of unknown functions that are present in or alongside an

operon can perform a related function. This family of modules will, therefore, include operon mapping tools, such as OperonMapper (Taboada et al., 2012).

- Pathways: Pathway information can improve the information on a protein by determining what processes it is involved in. Additional modules could potentially map a protein to the KEGG database or use predictors to close gaps in known pathways.
- Localization: The modules in the localization family will determine the localization of the protein in the cell and in which compartments it is expressed. This could be achieved by tools that predict the presence of a signal peptide, a GO-based predictor, and other specific methods such as PSORT (Horton et al., 2007).

These families would be associated with updated versions of the current databases and integrated with results from biological experiments (Figure 16). The gaps between these information families can be closed using experiments. Gene fusion and fission experiments could benefit the Interactions and Genome families, while phylogenetic profiling, GFP-fused expression, and Yeast 2 Hybrids (Y2H) could improve the sensitivity in the Genome and Interaction families respectively. A major challenge would be to develop a unified language that would integrate biological information and in-silico predictions, as well as novel predictors that aren't currently available.

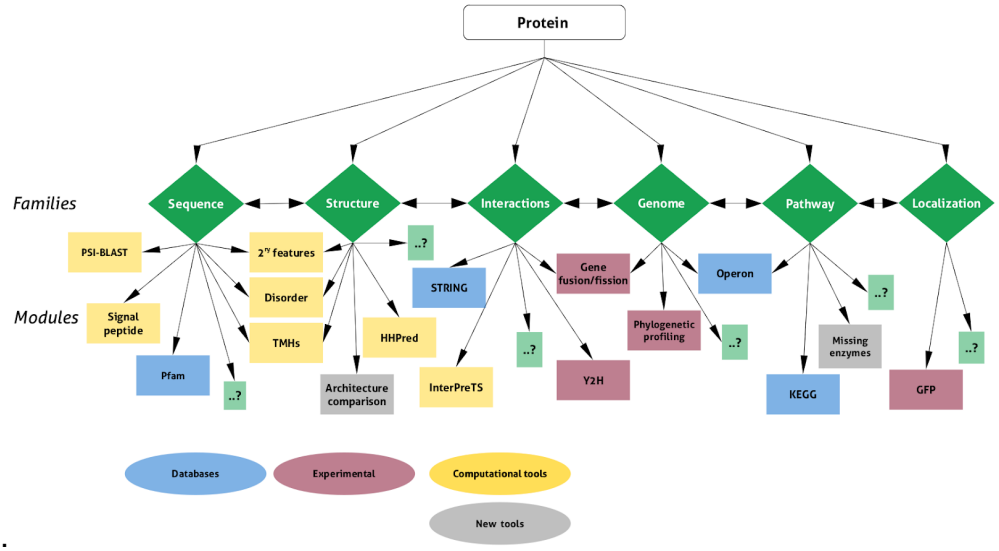


Figure 16. ICB architecture with its families, modules and tools (Courtesy of J.C. Gonzaléz Sanchez).

ICB validation

The ICB approach should be validated on model organisms with curated annotations and support of wet-lab experiments. The application on the PVC bacteria superphylum ameliorated its poor annotation status, but the ultimate test for the pipeline would be to reannotate from scratch *S. cerevisiae*, *E. coli* and other well-characterized organisms. The tool is valid if the same annotation status can be achieved and eventually improved by adding information on GOs or other features. A subsequent test would be to participate in CAFA and note how well the pipeline performs compared to other tools being developed at the current time.

Machine Learning

A promising approach to increase the sensitivity of ICB would be to use a Machine Learning layer to classify the GOs terms associated manually by SwissProt curators to a specific protein family and then use this information to create a consensus decision on the final protein identity. While automatic annotation usually relies on a single predictor,

manual curation assigns a function to a protein by looking at a combination of data. If we can train a neural network to recognize that a specific combination corresponds to a specific protein function, it would help in reaching the ultimate goal of ICB, which is to automate with confidence the slow process of manual curation.

Protein Architecture Detection in Distant Homologs

The architecture of a protein, defined as the presence of structural domains and their order in the protein, is more conserved than structure and sequence. Most of the available tools for protein annotation are based on sequence and structure homology, but an architecture-based system isn't currently available. Detecting remote ancestors or proteins shared by organisms that are far apart in the Tree of Life is the foundation of protein function prediction and identifying distant homologs could be helpful in understanding the evolution of a protein, its conservation, and its role in the cell biology. Previous studies based on architecture showed how some bacterial proteins in the PVCs have homologs only in Eukaryotes and this discovery was what kickstarted research on PVCs (Santarella-Mellwig R et al., 2010). We are currently developing HOUNDS, a tool for remote homologs detection based on architecture. Early results showed that some peroxisomal proteins in distant species can be found only using architecture-based methods, due to the low sequence similarity (as low as 5% for *Saccharomyces cerevisiae* and *Homo sapiens*, preliminary data).

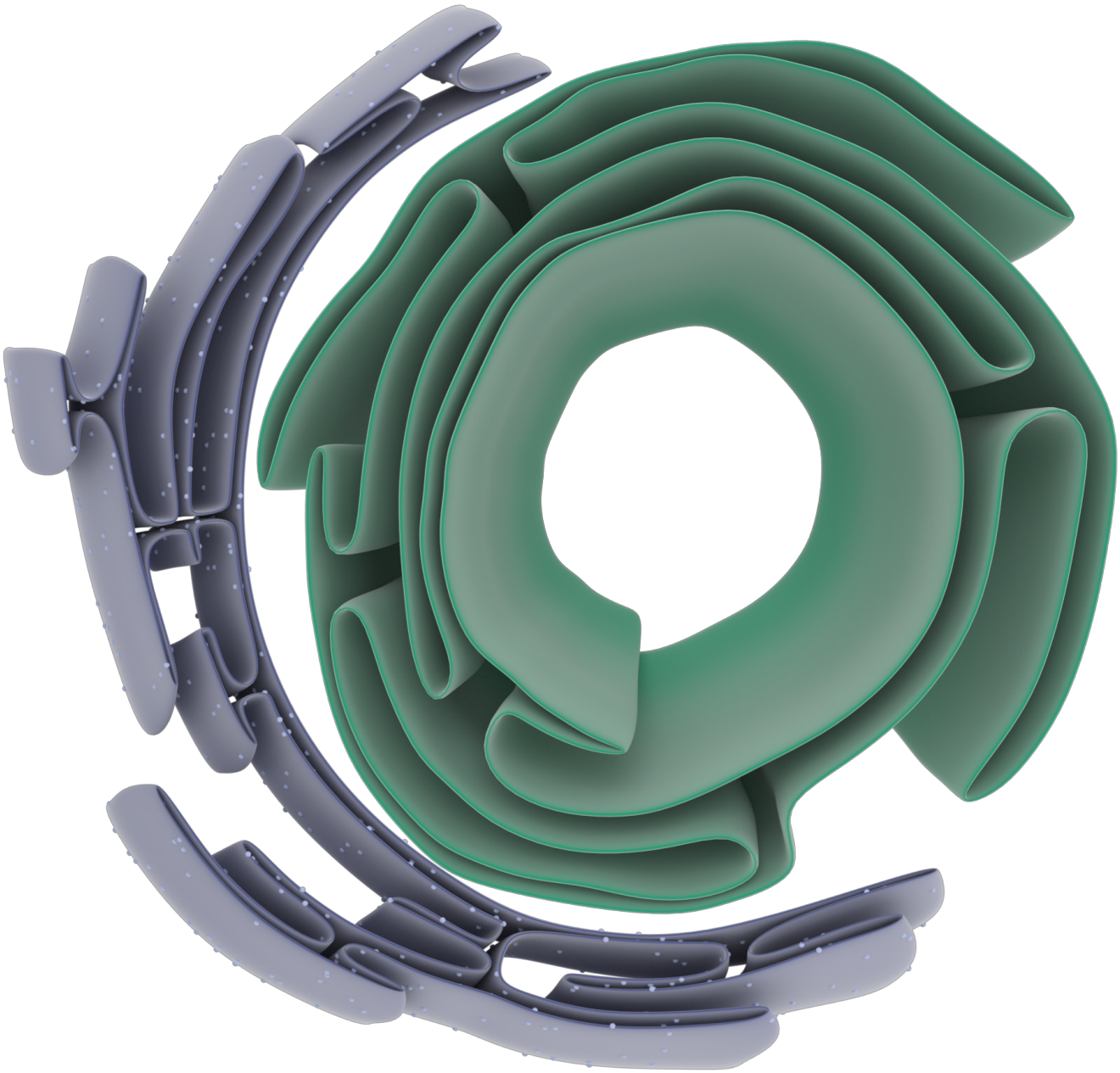
ICB Web-Server

As an addition to the ICB suite of computational tools, we planned to create an ICB webserver for on-demand analyses with a web interface. This allows the user to test the platform as seen on PVCbase, analyze a small batch of proteins without downloading the entire pipeline, and obtain some results without any previous bioinformatics knowledge.

ICB for the Dark Proteome

The ICB pipeline is being used to characterize a new recalculation of the Dark Proteome, focusing on assigning a function to the “unknown unknowns” proteins that are currently without a functional and structural assignment.

6 Conclusions



Conclusions

1 An integrative approach to protein function prediction generates a more confident protein identity.

2 An Integrative Cell Biology Pipeline was created, and it has been shown to vastly improve the knowledge of poorly characterized organisms.

3 In the case of PVC bacteria, the amount of “uncharacterized” proteins decreased from 46% to 25.5%.

4 Analyses of the data discovered an intermediate level of protein disorder in *Planctomycetes* when compared to other Bacteria and Eukaryotes.

5 ICB was used to annotate three macroalgal associated *Planctomycetes*, characterizing the features that distinguish them from PVCs living in other environments. Further analyses on proteins and metabolic pathways helped in elucidating their complex lifestyle at the interface with the algae.

6 PVCbase was created to collect and display the results obtained through ICB, alongside different tools that allow further analyses of the PVC bacteria superphylum.

7 A containerized version of ICB (ICBdocker) allows easy deployment of the system, in order for research groups focused on different organisms to obtain a more accurate proteome characterization with web-pages to browse the results.

References

Acehan, D., Santarella-Mellwig, R., & Devos, D. P. (2014). A bacterial tubulovesicular network. *Journal of Cell Science*, *127*(2), 277–280. https://doi.org/10.1242/jcs.137596

Aloy, P., & Russell, R. B. (2003). InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics (Oxford, England)*, *19*(1), 161–162. https://doi.org/https://doi.org/10.1093/bioinformatics/19.1.161

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/https://doi.org/10.1093/nar/25.17.3389

Babbitt, B. P. C., & Gerlt, J. A. (2001). NEW FUNCTIONS FROM OLD SCAFFOLDS : HOW NATURE REENGINEERS ENZYMES FOR NEW FUNCTIONS sign ,'' focuses on the recent applications of new methods in high. *Advances*, *55*, 1–28. https://doi.org/https://doi.org/10.1016/S0065-3233(01)55001-9

Boedeker, C., Schüler, M., Reintjes, G., Jeske, O., Van Teeseling, M. C. F., Jogler, M., … Jogler, C. (2017). Determining the bacterial cell biology of Planctomycetes. *Nature Communications*, *8*, 14853. https://doi.org/10.1038/ncomms14853

Bordin, N., & Devos, D. P. (2018). ICBdocker: a Docker image for proteome annotation and visualization. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty493

Bordin, N., González-Sánchez, J. C., & Devos, D. P. (2018). PVCbase: an integrated web resource for the PVC bacterial proteomes. *Database*, *2018*. https://doi.org/10.1093/database/bay042

Busch, D. J., Houser, J. R., Hayden, C. C., Sherman, M. B., Lafer, E. M., & Stachowiak, J. C. (2015). Intrinsically disordered proteins drive membrane curvature. *Nature Communications*, *6*(1), 7875. https://doi.org/10.1038/ncomms8875

Calhoun, S., Korczynska, M., Wichelecki, D. J., San Francisco, B., Zhao, S., Rodionov, D. A., … Sali, A. (2018). Prediction of enzymatic pathways by integrative pathway mapping. *ELife*, *7*. https://doi.org/10.7554/eLife.31097

Challis, R. J., Kumar, S., Stevens, L., & Blaxter, M. (2017). GenomeHubs: simple containerized setup of a custom Ensembl database and web server for any species. *Database*, *2017*. https://doi.org/10.1093/database/bax039

Cho, J.-C., Vergin, K. L., Morris, R. M., & Giovannoni, S. J. (2004). Lentisphaera araneosa gen. nov., sp. nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. *Environmental Microbiology*, *6*(6), 611–621. https://doi.org/10.1111/j.1462-2920.2004.00614.x

Devos, D. P., & Reynaud, E. G. (2010). Intermediate Steps. *Science*, *330*(6008), 1187–1188. https://doi.org/10.1126/science.1196720

Devos, D. P. (2014, January). PVC bacteria: Variation of, but not exception to, the Gram-negative cell plan. *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2013.10.008

Devos, D. P., & Ward, N. L. (2014). Mind the PVCs. *Environmental Microbiology*, *16*(5), 1217–1221. https://doi.org/10.1111/1462-2920.12349

Devos, D. P. (2013). Gemmata obscuriglobus. *Current Biology*, *23*(17), R705–R707. https://doi.org/10.1016/j.cub.2013.07.013

Devos, D. P. (2014). Re-interpretation of the evidence for the PVC cell plan supports a Gram-negative origin. *Antonie van Leeuwenhoek*, *105*(2), 271–274. https://doi.org/10.1007/s10482-013-0087-y

Devos, D., & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Structure, Function and Genetics*, *41*(1), 98–107. https://doi.org/10.1002/1097-0134(20001001)41:1<98::AID-PROT120>3.0.CO;2-S

Domazet-Lošo, T., Brajković, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, *23*(11), 533–539. https://doi.org/10.1016/j.tig.2007.08.014

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)*, *21*(16), 3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Earnshaw, W. C. (2013). Deducing protein function by forensic integrative cell biology. *PLoS Biology*, *11*(12), e1001742. https://doi.org/10.1371/journal.pbio.1001742

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Erdin, S., Lisewski, A. M., & Lichtarge, O. (2011). Protein function prediction: towards integration of similarity metrics. *Current Opinion in Structural Biology*, *21*(2), 180–188. https://doi.org/10.1016/J.SBI.2011.02.001

Faria, M., Bordin, N., Kizina, J., Harder, J., Devos, D., & Lage, O. M. (2017). Planctomycetes attached to algal surfaces: Insight into their genomes. *Genomics*. https://doi.org/10.1016/J.YGENO.2017.10.007

Faria, M., Bordin, N., Kizina, J., Harder, J., Devos, D., & Lage, O. M. (2017). Planctomycetes attached to algal surfaces: Insight into their genomes. *Genomics*. https://doi.org/10.1016/j.ygeno.2017.10.007

Fieseler, L., Horn, M., Wagner, M., & Hentschel, U. (2004). Discovery of the novel candidate phylum Poribacteria in marine sponges. *Applied and Environmental Microbiology*, *70*(6), 3724–3732. https://doi.org/10.1128/AEM.70.6.3724-3732.2004

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., … Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512. https://doi.org/10.1126/science.7542800

FRANKLIN, R. E., & GOSLING, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, *171*(4356), 740–741. https://doi.org/10.1038/171740a0

Fuerst, J. A. (2013). The PVC superphylum: exceptions to the bacterial definition? *Antonie van Leeuwenhoek*, *104*(4), 451–466. https://doi.org/10.1007/s10482-013-9986-1

Fuerst, J. A., & Sagulenko, E. (2014). Towards understanding the molecular mechanism of the endocytosis-like process in the bacterium Gemmata obscuriglobus. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1843*(8), 1732–1738. https://doi.org/10.1016/J.BBAMCR.2013.10.002

Galperin, M. Y., & Koonin, E. V. (2012). Divergence and Convergence in Enzyme Evolution. *Journal of Biological Chemistry*, *287*(1), 21–28. https://doi.org/10.1074/jbc.R111.241976

Glöckner, F. O., Kube, M., Bauer, M., Teeling, H., Lombardot, T., Ludwig, W., … Reinhardt, R. (2003). Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proceedings of the National Academy of Sciences*, *100*(14), 8298–8303. https://doi.org/10.1073/pnas.1431443100

González-Sánchez, J. C., Costa, R., & Devos, D. P. (2015). A multi-functional tubulovesicular network as the ancestral eukaryotic endomembrane system. *Biology*, *4*(2), 264–281. https://doi.org/10.3390/biology4020264

Green, M. L., & Karp, P. D. (2007). Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics*, *23*(13), i205–i211. https://doi.org/10.1093/bioinformatics/btm213

Gupta, R. S., Bhandari, V., & Naushad, H. S. (2012). Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. *Frontiers in Microbiology*, *3*, 327. https://doi.org/10.3389/fmicb.2012.00327

Gutiérrez, F. I., Rodriguez-Valenzuela, F., Ibarra, I. L., Devos, D. P., & Melo, F. (2016). Efficient and automated large-scale detection of structural relationships in proteins with a flexible aligner. *BMC Bioinformatics*, *17*(1), 20. https://doi.org/10.1186/s12859-015-0866-8

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., & Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, *35*(Web Server), W585–W587. https://doi.org/10.1093/nar/gkm259

Huberts, D. H. E. W., & van der Klei, I. J. (2010). Moonlighting proteins: an intriguing mode of multitasking. *Biochimica et Biophysica Acta*, *1803*(4), 520–525. https://doi.org/10.1016/j.bbamcr.2010.01.022

Imai, M., Watanabe, T., Hatta, M., Das, S. C., Ozawa, M., Shinya, K., … Kawaoka, Y. (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, *486*(7403), 420–428. https://doi.org/10.1038/nature10831

Jeske, O., Schüler, M., Schumann, P., Schneider, A., Boedeker, C., Jogler, M., … Jogler, C. (2015). Planctomycetes do possess a peptidoglycan cell wall. *Nature Communications*, *6*(1), 7116. https://doi.org/10.1038/ncomms8116

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., … Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, *17*(1), 184. https://doi.org/10.1186/s13059-016-1037-6

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, *27*(6), 757–763. https://doi.org/10.1093/bioinformatics/btr010

Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., & Brunak, S. (2002). Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering, Design and Selection*, *15*(4), 287–296. https://doi.org/10.1093/protein/15.4.287

Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., & Church, G. M. (2006). Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, *7*(1), 177. https://doi.org/10.1186/1471-2105-7-177

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Lage, O. M., & Bondoso, J. (2011). Planctomycetes diversity associated with macroalgae. *FEMS Microbiology Ecology*, *78*(2), 366–375. https://doi.org/10.1111/j.1574-6941.2011.01168.x

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences*, *106*(27), 11079–11084. https://doi.org/10.1073/pnas.0905029106

Liechti, G. W., Kuru, E., Hall, E., Kalinda, A., Brun, Y. V., VanNieuwenhze, M., & Maurelli, A. T. (2014). A new metabolic cell-wall labelling method reveals peptidoglycan in Chlamydia trachomatis. *Nature*, *506*(7489), 507–510. https://doi.org/10.1038/nature12892

Lindsay, M. R., Webb, R. I., Strous, M., Jetten, M. S. M., Butler, M. K., Forde, R. J., & Fuerst, J. A. (2001). Cell compartmentalisation in planctomycetes: Novel types of structural organisation for the bacterial cell. *Archives of Microbiology*, *175*(6), 413–429. https://doi.org/10.1007/s002030100280

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., … Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biology*, *10*(2), 207. https://doi.org/10.1186/gb-2009-10-2-207

Lonhienne, T. G. A., Sagulenko, E., Webb, R. I., Lee, K.-C., Franke, J., Devos, D. P., … Fuerst, J. A. (2010). Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences*, *107*(29), 12883–12888. https://doi.org/10.1073/pnas.1001085107

Lonhienne, T. G. A., Sagulenko, E., Webb, R. I., Lee, K.-C., Franke, J., Devos, D. P., … Fuerst, J. A. (2010). Endocytosis-like protein uptake in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 12883–12888. https://doi.org/10.1073/pnas.1001085107

McDowall, J., & Hunter, S. (2011). InterPro Protein Classification (pp. 37–47). https://doi.org/10.1007/978-1-60761-977-2_3

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404–405. https://doi.org/10.1093/bioinformatics/16.4.404

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*. https://doi.org/10.1097/01.NND.0000320699.47006.a3

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics*, *23*(3), ii–iv. https://doi.org/10.1002/prot.340230303

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999, January 1). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/27.1.29

Omelchenko, M. V, Galperin, M. Y., Wolf, Y. I., & Koonin, E. V. (2010). Non-homologous isofunctional enzymes: A systematic analysis of alternative solutions in enzyme evolution. *Biology Direct*, *5*(1), 31. https://doi.org/10.1186/1745-6150-5-31

Ooi, H. S., Kwo, C. Y., Wildpaner, M., Sirota, F. L., Eisenhaber, B., Maurer-Stroh, S., … Schneider, G. (2009). ANNIE: integrated de novo protein sequence annotation. *Nucleic Acids Research*, *37*(Web Server issue), W435-40. https://doi.org/10.1093/nar/gkp254

Pearson, A., Budin, M., & Brocks, J. J. (2003). Phylogenetic and biochemical evidence for sterol synthesis in the bacterium Gemmata obscuriglobus. *Proceedings of the National Academy of Sciences*, *100*(26), 15352–15357. https://doi.org/10.1073/pnas.2536559100

Pearson, W. R. (2015). Protein Function Prediction: Problems and Pitfalls. In *Current Protocols in Bioinformatics* (Vol. 51, p. 4.12.1-4.12.8). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi0412s51

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., … O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(52), 15898–15903. https://doi.org/10.1073/pnas.1508380112

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Picco, A., Irastorza-Azcarate, I., Specht, T., Böke, D., Pazos, I., Rivier-Cordey, A.-S., … Gallego, O. (2017). The In Vivo Architecture of the Exocyst Provides Structural Basis for Exocytosis. *Cell*, *168*(3), 400–412.e18. https://doi.org/10.1016/j.cell.2017.01.004

Pietrosemoli, N., Pancsa, R., & Tompa, P. (2013). Structural disorder provides increased adaptability for vesicle trafficking pathways. *PLoS Computational Biology*, *9*(7), e1003144. https://doi.org/10.1371/journal.pcbi.1003144

Pilhofer, M., Aistleitner, K., Biboy, J., Gray, J., Kuru, E., Hall, E., … Jensen, G. J. (2013). Discovery of chlamydial peptidoglycan reveals bacteria with murein sacculi but without FtsZ. *Nature Communications*, *4*(1), 2856. https://doi.org/10.1038/ncomms3856

Pilhofer, M., Rappl, K., Eckl, C., Bauer, A. P., Ludwig, W., Schleifer, K.-H., & Petroni, G. (2008). Characterization and evolution of cell division and cell wall synthesis genes in the bacterial phyla Verrucomicrobia, Lentisphaerae, Chlamydiae, and Planctomycetes and phylogenetic comparison with rRNA genes. *Journal of Bacteriology*, *190*(9), 3192–3202. https://doi.org/10.1128/JB.01797-07

Piovesan, D., Luigi Martelli, P., Fariselli, P., Zauli, A., Rossi, I., & Casadio, R. (2011). BAR-PLUS: the Bologna Annotation Resource Plus for functional and structural annotation of protein sequences. *Nucleic Acids Research*, *39*(suppl), W197–W202. https://doi.org/10.1093/nar/gkr292

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., … Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, *10*(3), 221–227. https://doi.org/10.1038/nmeth.2340

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Reynaud, E. G., & Devos, D. P. (2011). Transitional forms between the three domains of life and evolutionary implications. *Proceedings. Biological Sciences*, *278*(1723), 3321–3328. https://doi.org/10.1098/rspb.2011.1581

Rivas-Marín, E., Canosa, I., & Devos, D. P. (2016). Evolutionary Cell Biology of Division Mode in the Bacterial Planctomycetes-Verrucomicrobia- Chlamydiae Superphylum. *Frontiers in Microbiology*, *7*, 1964. https://doi.org/10.3389/fmicb.2016.01964

Rivas-Marín, E., & Devos, D. P. (2018). The Paradigms They Are a-Changin': past, present and future of PVC bacteria research. *Antonie van Leeuwenhoek*, *111*(6), 785–799. https://doi.org/10.1007/s10482-017-0962-z

Šali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626

Santarella-Mellwig, R., Franke, J., Jaedicke, A., Gorjanacz, M., Bauer, U., Budd, A., … Devos, D. P. (2010). The Compartmentalized Bacteria of the Planctomycetes-Verrucomicrobia-Chlamydiae Superphylum Have Membrane Coat-Like Proteins. *PLoS Biology*, *8*(1), e1000281. https://doi.org/10.1371/journal.pbio.1000281

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Computational Biology*, *5*(12), e1000605. https://doi.org/10.1371/journal.pcbi.1000605

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Smith, A. A. T., Belda, E., Viari, A., Medigue, C., & Vallenet, D. (2012). The CanOE Strategy: Integrating Genomic and Metabolic Contexts across Multiple Prokaryote Genomes to Find Candidate Genes for Orphan Enzymes. *PLoS Computational Biology*, *8*(5), e1002540. https://doi.org/10.1371/journal.pcbi.1002540

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)*, *21*(7), 951–960. https://doi.org/10.1093/bioinformatics/bti125

Strous, M., Kuenen, J. G., & Jetten, M. S. M. (1999). Key physiology of anaerobic ammonium oxidation. *Applied and Environmental Microbiology*, *65*(7), 3248–3250. https://doi.org/10.1.1.318.4408

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., … von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, *43*(Database issue), D447-52. https://doi.org/10.1093/nar/gku1003

Taboada, B., Ciria, R., Martinez-Guerrero, C. E., & Merino, E. (2012). ProOpDB: Prokaryotic Operon DataBase. *Nucleic Acids Research*, *40*(D1), D627–D631. https://doi.org/10.1093/nar/gkr1020

Tantos, A., Han, K.-H., & Tompa, P. (2012). Intrinsic disorder in cell signaling and gene transcription. *Molecular and Cellular Endocrinology*, *348*(2), 457–465. https://doi.org/10.1016/j.mce.2011.07.015

Tatusova, T., Ciufo, S., Federhen, S., Fedorov, B., McVeigh, R., O'Neill, K., … Zaslavsky, L. (2015). Update on RefSeq microbial genomes resources. *Nucleic Acids Research*, *43*(D1), D599–D605. https://doi.org/10.1093/nar/gku1062

Tiwari, A. K., & Srivastava, R. (2014). A survey of computational intelligence techniques in protein function prediction. *International Journal of Proteomics*, *2014*, 845479. https://doi.org/10.1155/2014/845479

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(D1), D204–D212. https://doi.org/10.1093/nar/gku989

Van Niftrik, L. A., Fuerst, J. A., Sinninghe Damsté, J. S., Kuenen, J. G., Jetten, M. S. M., & Strous, M. (2004). The anammoxosome: An intracytoplasmic compartment in anammox bacteria. *FEMS Microbiology Letters*, *233*(1), 7–13. https://doi.org/10.1016/j.femsle.2004.01.044

van Niftrik, L., & Jetten, M. S. M. (2012). Anaerobic ammonium-oxidizing bacteria: unique microorganisms with exceptional properties. *Microbiology and Molecular Biology Reviews : MMBR*, *76*(3), 585–596. https://doi.org/10.1128/MMBR.05025-11

van Teeseling, M. C. F., Mesman, R. J., Kuru, E., Espaillat, A., Cava, F., Brun, Y. V, … van Niftrik, L. (2015). Anammox Planctomycetes have a peptidoglycan cell wall. *Nature Communications*, *6*(1), 6878. https://doi.org/10.1038/ncomms7878

Wagner, M., & Horn, M. (2006). The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Current Opinion in Biotechnology*, *17*(3), 241–249. https://doi.org/10.1016/j.copbio.2006.05.005

Watson, J., & Crick, F. (1953). Molecular structure of nucleic acids: A Structure for Deoxyribose Nucleic Acid. *Nature.*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Wilkins, M. H. F., Stokes, A. R., & Wilson, H. R. (1953). Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature*, *171*(4356), 738–740. Retrieved from http://www.citeulike.org/group/7862/article/3804333

Yamada, T., Waller, A. S., Raes, J., Zelezniak, A., Perchat, N., Perret, A., … Bork, P. (2012). Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. *Molecular Systems Biology*, *8*, 581. https://doi.org/10.1038/msb.2012.13

Yamanishi, Y., Mihara, H., Osaki, M., Muramatsu, H., Esaki, N., Sato, T., … Kanehisa, M. (2007). Prediction of missing enzyme genes in a bacterial metabolic network. *FEBS Journal*, *274*(9), 2262–2273. https://doi.org/10.1111/j.1742-4658.2007.05763.x

Ye, Y., Osterman, A., Overbeek, R., & Godzik, A. (2005). Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, *21*(Suppl 1), i478–i486. https://doi.org/10.1093/bioinformatics/bti1052

Yee, B., Sagulenko, E., Morgan, G. P., Webb, R. I., & Fuerst, J. A. (2012). Electron tomography of the nucleoid of Gemmata obscuriglobus reveals complex liquid crystalline cholesteric structure. *Frontiers in Microbiology*, *3*, 326. https://doi.org/10.3389/fmicb.2012.00326

Zhu, J., Sova, P., Xu, Q., Dombek, K. M., Xu, E. Y., Vu, H., … Schadt, E. E. (2012). Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLoS Biology*, *10*(4), e1001301. https://doi.org/10.1371/journal.pbio.1001301